High-Confidence Attack Detection via Wasserstein-Metric Computations

Dan Li, *Graduate Student Member, IEEE*, and Sonia Martínez, *Fellow, IEEE*

***Abstract*—This letter considers a sensor attack and fault detection problem for linear cyber-physical systems, which are subject to system noise that can obey an unknown light-tailed distribution. We propose a new threshold-based detection mechanism that employs the Wasserstein metric, and which guarantees system performance with high confidence with a finite number of measurements. The proposed detector may generate false alarms with a rate Δ in normal operation, where Δ can be tuned to be arbitrarily small by means of a *benchmark distribution*. Thus, the proposed detector is sensitive to sensor attacks and faults which have a statistical behavior that is different from that of the system noise. We quantify the impact of *stealthy* attacks on open-loop stable systems—which perturb the system operation while producing false alarms consistent with the natural system noise—via a *probabilistic* reachable set. Tractable implementation is enabled via a linear optimization to compute the detection measure and a semidefinite program to bound the reachable set.**

***Index Terms*—Networked control systems, fault detection.**

I. Introduction

Cyber-physical Systems (CPS) are physical processes that are tightly integrated with computation and communication systems for monitoring and control. These systems are usually complex, large-scale and insufficiently supervised, making them vulnerable to attacks [1], [2]. A significant literature has studied various *denial of service* [3], *false data-injection* [4], [5], *replay* [6], [7], *sensor, and integrity* attacks [8]–[11] in a control-theoretic framework, by comparing estimation and measurements w.r.t. predefined metrics. However, attacks could be *stealthy*, and exploit knowledge of the system structure, uncertainty and noise information to inflict significant damage on the physical system while avoiding detection. This motivates the characterization of the impact of stealthy attacks via e.g., reachability set analysis [9], [12], [13]. To ensure computational tractability, these works assume either Gaussian or bounded system noise. However, these assumptions fall short in modeling all natural disturbances that can affect a system. When designing detectors, an added difficulty is in obtaining tractable computations that can handle these more general distributions. More recently, novel measure of concentration has opened the way for online tractable and robust attack detection with probability guarantees under uncertainty. A first attempt in this direction is [14], which exploits the Chebyshev's inequality to design a detector, and characterizes stealthy attacks on stable systems affected by bounded system noises. With the aim of obtaining a less conservative detection mechanism, we leverage an alternative measure-concentration result via Wasserstein metric. This metric is built from data gathered on the system, and can provide significantly sharper results than those stemming from the Chebyshev inequality. In particular, we address the following question for linear CPSs: *How to design an online attack-detection mechanism that is robust to light-tailed distributions of system noise while remaining sensitive to attacks and limiting the impact of the stealthy attack?*

Statement of Contributions: To address the previous question: 1) We propose a novel detection measure, which employs the Wasserstein distance between the benchmark and a distribution of the residual sequence obtained online. 2) We propose a novel threshold-detection mechanism, which exploits measure-of-concentration results to guarantee the robust detection of an attack with high confidence using a finite set of data, and which further enables the robust tuning of the false alarm rate. The proposed detector can effectively identify real-time attacks when its behavior differs from that of the system noise. In addition, the detector can handle systems noises that are not necessarily distributed as Gaussian. 3) We propose a quantifiable, probabilistic state-reachable set, which reveals the impact of the stealthy attacker and system noise on open loop stable systems with high probability. 4) To implement the proposed mechanism, we formulate a linear optimization problem and a semidefinite problem for the computation of the detection measure as well as the reachable set, respectively. We encourage the reader to access online version [15] for the complete development of the ideas.

II. CPS and Its Normal Operation

A remotely-observed, cyber-physical system subject to sensor-measurement attacks, is described as a discrete-time, stochastic, linear, and time-invariant system

$$\begin{aligned} \boldsymbol{x}(t+1) &= A\boldsymbol{x}(t) + B\boldsymbol{u}(t) + \boldsymbol{w}(t), \\ \boldsymbol{y}(t) &= C\boldsymbol{x}(t) + \boldsymbol{v}(t) + \boldsymbol{\gamma}(t), \end{aligned} \tag{1}$$

Manuscript received March 17, 2020; revised May 21, 2020; accepted June 2, 2020. Date of publication June 16, 2020; date of current version June 30, 2020. This work was supported in part by Office of Naval Research under Grant N00014-19-1-2471, and in part by Air Force Office of Scientific Research under Grant FA9550-19-1-0235. Recommended by Senior Editor V. Ugrinovskii. *(Corresponding author: Dan Li.)*

The authors are with the Mechanical and Aerospace Engineering, University of California at San Diego, La Jolla, CA 92092 USA (e-mail: lidan@ucsd.edu; soniamd@ucsd.edu).

Digital Object Identifier 10.1109/LCSYS.2020.3002689

2475-1456 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on October 24,2023 at 20:05:42 UTC from IEEE Xplore. Restrictions apply.

where $\boldsymbol{x}(t) \in \mathbb{R}^n$, $\boldsymbol{u}(t) \in \mathbb{R}^m$ and $\boldsymbol{y}(t) \in \mathbb{R}^p$ denote the system state, input and output at time $t \in \mathbb{N}$, respectively. The state matrix A, input matrix B and output matrix C are assumed to be known in advance. In particular, we assume that the pair (A, B) is stabilizable, and (A, C) is detectable. The process noise $\boldsymbol{w}(t) \in \mathbb{R}^n$ and output noise $\boldsymbol{v}(t) \in \mathbb{R}^p$ are independent zero-mean random vectors. We assume that each $\boldsymbol{w}(t)$ and $\boldsymbol{v}(t)$ are independent and identically distributed (i.i.d.) over time. We denote their (unknown, not-necessarily equal) distributions by $\mathbb{P}_{\boldsymbol{w}}$ and $\mathbb{P}_{\boldsymbol{v}}$, respectively. In addition, we assume that $\mathbb{P}_{\boldsymbol{w}}$ and $\mathbb{P}_{\boldsymbol{v}}$ are light-tailed,[1] excluding scenarios of systems operating under extreme events, or subject to large delays. In fact, Gaussian, Sub-Gaussian, Exponential distributions, and any distribution with a compact support set are admissible. This distribution class is sufficient to characterize the uncertainty or noise of many practical problems.

An additive sensor-measurement attack is implemented via $\boldsymbol{\gamma}(t) \in \mathbb{R}^p$ in (1), on which we assume the following

Assumption 1 (Attack Model): It holds that 1) $\boldsymbol{\gamma}(t) = \mathbf{0}$ whenever there is no attack; 2) the attacker can modulate any component of $\boldsymbol{\gamma}(t)$ at any time; 3) the attacker has unlimited computational resources and access to system information, e.g., A, B, C, $\boldsymbol{u}$, $\mathbb{P}_{\boldsymbol{w}}$ and $\mathbb{P}_{\boldsymbol{v}}$ to decide on $\boldsymbol{\gamma}(t)$, $t \in \mathbb{N}$.

A. Normal System Operation

Here, we introduce the state observer that enables prediction in the absence of attacks ($\boldsymbol{\gamma}(t) = \mathbf{0}$). Since the distribution of system noise is unknown, we identify a benchmark distribution to capture this unknown distribution with high confidence.

To predict the system behavior, we employ a Kalman filter

$$\begin{aligned} \hat{\boldsymbol{x}}(t+1) &= A\hat{\boldsymbol{x}}(t) + B\boldsymbol{u}(t) + L(t)\big(\boldsymbol{y}(t) - \hat{\boldsymbol{y}}(t)\big), \\ \hat{\boldsymbol{y}}(t) &= C\hat{\boldsymbol{x}}(t), \end{aligned}$$

where $\hat{\boldsymbol{x}}(t)$ is the state estimate and $L(t) \equiv L$ is the steady-state Kalman gain matrix. As the pair (A, C) is detectable, the gain L is selected to bring the eigenvalues of $A - LC$ inside the unit circle. This ensures that the estimation error $\boldsymbol{e}(t) := \boldsymbol{x}(t) - \hat{\boldsymbol{x}}(t)$ satisfies

$$\mathbb{E}[\boldsymbol{e}(t)] \to 0 \text{ as } t \to \infty, \quad \text{for any } \boldsymbol{x}(0),\ \hat{\boldsymbol{x}}(0).$$

We additionally consider the estimated state feedback $\boldsymbol{u}(t) = K\hat{\boldsymbol{x}}(t)$, where K is selected to make the next system stable[2]

$$\boldsymbol{\xi}(t+1) = F\boldsymbol{\xi}(t) + G\boldsymbol{\sigma}(t), \tag{2}$$

where $\boldsymbol{\xi}(t) := (\boldsymbol{x}(t), \boldsymbol{e}(t))^\top$, $\boldsymbol{\sigma}(t) := (\boldsymbol{w}(t), \boldsymbol{v}(t) + \boldsymbol{\gamma}(t))^\top$,

$$F = \begin{bmatrix} A + BK & -BK \\ 0 & A - LC \end{bmatrix}, G = \begin{bmatrix} I & 0 \\ I & -L \end{bmatrix} \text{ and some } \boldsymbol{\xi}(0).$$

Consider the system initially operates normally after selecting L and K, and assume that the augmented system (2) is in steady state, i.e., $\mathbb{E}[\boldsymbol{\xi}(t)] = \mathbf{0}$. In order to design our attack detector, we need a characterization of the distribution of the *residue* $\boldsymbol{r}(t) \in \mathbb{R}^p$, evaluating the difference between what we measure and what we expect to receive:

$$\boldsymbol{r}(t) := \boldsymbol{y}(t) - \hat{\boldsymbol{y}}(t) = C\boldsymbol{e}(t) + \boldsymbol{v}(t) + \boldsymbol{\gamma}(t).$$

When there is no attack, it can be verified that $\boldsymbol{r}(t)$ is zero-mean, and light-tailed.[3] Let us denote its unknown distribution by $\mathbb{P}_{\boldsymbol{r}}$. We assume that a finite, but large number N of i.i.d. samples of $\mathbb{P}_{\boldsymbol{r}}$, are accessible, and acquired by collecting $\boldsymbol{r}(t)$ for a sufficiently large time. We call these i.i.d. samples a *benchmark data set*, $\Xi_{\mathrm{B}} := \{\boldsymbol{r}^{(i)} = \boldsymbol{y}^{(i)} - \hat{\boldsymbol{y}}^{(i)}\}_{i=1}^N$, and construct the resulting empirical distribution $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ by

$$\mathbb{P}_{\boldsymbol{r},\mathrm{B}} := \frac{1}{N}\sum_{i=1}^{N} \delta_{\{\boldsymbol{r}^{(i)}\}},$$

where the operator δ is the mass function, and the subscript B indicates that $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ is the benchmark distribution of the data. We claim that $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ provides a characterization of the effect of the noise on (2) via the following result:

Theorem 1 (Measure of Concentration [17, Application of Th. 2]): If $\mathbb{P}_{\boldsymbol{r}}$ is a q-light-tailed distribution for some $q \geq 1$, then for a given $\beta \in (0, 1)$, the following holds

$$\mathrm{Prob}\big(d_{W,q}(\mathbb{P}_{\boldsymbol{r},\mathrm{B}}, \mathbb{P}_{\boldsymbol{r}}) \leq \epsilon_{\mathrm{B}}\big) \geq 1 - \beta,$$

where Prob denotes the Probability of the samples in $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$, $d_{W,q}$ denotes the q-Wasserstein metric,[4] and the parameter ϵ_{B} is selected as

$$\epsilon_{\mathrm{B}} := \begin{cases} \left(\frac{\log(c_1\beta^{-1})}{c_2 N}\right)^{q/a}, & \text{if } N < \frac{\log(c_1\beta^{-1})}{c_2}, \\ \bar{\epsilon}, & \text{if } N \geq \frac{\log(c_1\beta^{-1})}{c_2}, \end{cases} \tag{3}$$

for some constant[5] $a > q$, c_1, $c_2 > 0$, and $\bar{\epsilon}$ is such that $c_2 N(\bar{\epsilon})^{\max\{2,p/q\}} = \log(c_1\beta^{-1})$, if $p \neq 2q$, or $\frac{\bar{\epsilon}}{\log(2+1/\bar{\epsilon})} = (\frac{\log(c_1\beta^{-1})}{c_2 N})^{1/2}$, if $p = 2q$, where p is the dimension of $\boldsymbol{r}$.

Theorem 1 provides a probabilistic bound ϵ_{B} on the q-Wasserstein distance between $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ and $\mathbb{P}_{\boldsymbol{r}}$, with a confidence at least $1 - \beta$. It indicates how to tune the parameter β and the number of benchmark samples N that are needed to find a sufficiently good approximation of $\mathbb{P}_{\boldsymbol{r}}$, by means of $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$. In this way, given an ϵ, we can increase our confidence $(1 - \beta)$ on whether $\mathbb{P}_{\boldsymbol{r}}$ and $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ are within distance ϵ, by increasing the number of samples. We assume that $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ has been determined in advance, selecting a very large (unique) N to ensure various very small bounds ϵ_{B} associated with various β.

III. Threshold-Based Robust Detection of Attacks, and Stealthiness

This section presents our online detection procedure, and a threshold-based detector with high-confidence performance

[1] For a random vector $\boldsymbol{w}$ such that $\boldsymbol{w} \sim \mathbb{P}$, we say $\mathbb{P}$ is q-light-tailed, $q = 1, 2, \ldots$, if $c := \mathbb{E}_{\mathbb{P}}[\exp(b\|\boldsymbol{w}\|^a)] < \infty$ for some $a > q$ and $b > 0$. All examples listed have a moment generating function, so their exponential moment can be constructed for at least $q = 1$.

[2] System (2) is input-to-state stable in probability (ISSp) relative to any compact set $\mathcal{A}$ which contains the origin, if we select K such that eigenvalues of the matrix $A + BK$ are inside the unit circle, see [16].

[3] This can be checked from the definition in footnote 1, and follows from $\boldsymbol{r}(t)$ being a linear combination of zero-mean q-light-tailed distributions.

[4] Let $\mathcal{M}_q(\mathcal{Z})$ denote the space of all q-light-tailed probability distributions supported on $\mathcal{Z} \subset \mathbb{R}^p$. Then for any two distributions $\mathbb{Q}_1$, $\mathbb{Q}_2 \in \mathcal{M}_q(\mathcal{Z})$, the q-Wasserstein metric [18] $d_{W,q}{:}\mathcal{M}_q(\mathcal{Z}) \times \mathcal{M}_q(\mathcal{Z}) \to \mathbb{R}_{\geq 0}$ is defined by $d_{W,q}(\mathbb{Q}_1, \mathbb{Q}_2) := (\min_{\Pi} \int_{\mathcal{Z}\times\mathcal{Z}} \ell^q(\xi_1, \xi_2)\Pi(d\xi_1, d\xi_2))^{1/q}$, where Π is in a set of all the probability distributions on $\mathcal{Z} \times \mathcal{Z}$ with marginals $\mathbb{Q}_1$ and $\mathbb{Q}_2$. The cost $\ell(\xi_1, \xi_2) := \|\xi_1 - \xi_2\|$ is a norm on $\mathcal{Z}$.

[5] The parameter a is determined as in the definition of $\mathbb{P}_{\boldsymbol{r}}$ and the constants c_1, c_2 depend on q, m, and $\mathbb{P}_{\boldsymbol{r}}$ (via a, b, c). When information on $\mathbb{P}_{\boldsymbol{r}}$ is absent, the parameters a, c_1 and c_2 can be determined in a data-driven fashion using sufficiently many samples of $\mathbb{P}_{\boldsymbol{r}}$. See [17] for details.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on October 24,2023 at 20:05:42 UTC from IEEE Xplore. Restrictions apply.

guarantees. Then, we propose a tractable computation of the detection measure used for online detection. We finish the section by introducing a class of stealthy attacks.

Online Detection Procedure (ODP): At each time $t \geq T$, we construct a T-step detector distribution

$$\mathbb{P}_{\boldsymbol{r},\mathrm{D}} := \frac{1}{T}\sum_{j=0}^{T-1}\delta_{\{\boldsymbol{r}(t-j)\}},$$

where $\boldsymbol{r}(t-j)$ is the residue data collected independently at time $t-j$, for $j \in \{0,\dots,T-1\}$. Then with a given q and a threshold $\alpha > 0$, we consider the *detection measure*

$$z(t) := d_{W,q}(\mathbb{P}_{\boldsymbol{r},\mathrm{B}}, \mathbb{P}_{\boldsymbol{r},\mathrm{D}}), \tag{4}$$

and the *attack detector*

$$\begin{cases} z(t) \leq \alpha, & \text{no alarm at } t: \ \mathrm{Alarm}(t) = 0, \\ z(t) > \alpha, & \text{alarm at } t: \ \mathrm{Alarm}(t) = 1, \end{cases} \tag{5}$$

with Alarm(t) the sequence of alarms generated online based on the previous threshold. The distribution $\mathbb{P}_{\boldsymbol{r},\mathrm{D}}$ uses a small number T of samples to ensure the online computational tractability of $z(t)$, so $\mathbb{P}_{\boldsymbol{r},\mathrm{D}}$ is highly dependent on instantaneous samples. Thus, $\mathbb{P}_{\boldsymbol{r},\mathrm{D}}$ may significantly deviate from the true $\mathbb{P}_{\boldsymbol{r}}$, and from $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$. Therefore, even if there is no attack, the attack detector is expected to generate false alarms due to the system noise as well as an improper selection of the threshold α. In the following, we discuss how to select an α that is robust to the system noise and which results in a desired false alarm rate. Note that the value α should be small to be able to distinguish attacks from noise, as discussed later.

Lemma 1 (Selection of α for Robust Detectors): Given parameters N, T, q, β, and a desired false alarm rate $\Delta > \beta$ at time t, if we select the threshold α as

$$\alpha := \epsilon_{\mathrm{B}} + \epsilon_{\mathrm{D}},$$

where ϵ_{B} is chosen as in (3) and ϵ_{D} is selected following the ϵ_{B}-formula (3), but with T and $\frac{\Delta-\beta}{1-\beta}$ in place of N and β, respectively. Then, the detection measure (4) satisfies

$$\mathrm{Prob}(z(t) \leq \alpha) \geq 1 - \Delta,$$

for any zero-mean q-light-tailed underlying distribution $\mathbb{P}_{\boldsymbol{r}}$.

Proof: To prove this, $z(t) \leq d_{W,q}(\mathbb{P}_{\boldsymbol{r},\mathrm{B}}, \mathbb{P}_{\boldsymbol{r}}) + d_{W,q}(\mathbb{P}_{\boldsymbol{r},\mathrm{D}}, \mathbb{P}_{\boldsymbol{r}})$ follows from the triangular inequality. Then we apply Theorem 1 for each $d_{W,q}$ term, and the fact that $\mathrm{Prob}(d_{W,q}(\mathbb{P}_{\boldsymbol{r},\mathrm{D}}, \mathbb{P}_{\boldsymbol{r}}) \leq \epsilon_{\mathrm{D}}) \geq 1 - \frac{\Delta-\beta}{1-\beta}$. Note also that samples of $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ and $\mathbb{P}_{\boldsymbol{r},\mathrm{D}}$ are collected independently. ■

Lemma 1 ensures that the false alarm rate is no higher than Δ when there is no attack, i.e.,

$$\mathrm{Prob}(\mathrm{Alarm}(t) = 1 \mid \text{no attack}) \leq \Delta, \quad \forall\, t.$$

Note that the rate Δ can be selected by properly choosing the threshold α. Intuitively, if we fix all the other parameters, then the smaller the rate Δ, the larger the threshold α. Also, large values of N, T, $1-\beta$ contribute to small α.

Computation of detection measure: To achieve a tractable computation of the detection measure $z(t)$, we leverage the definition of the Wasserstein distance (see footnote 4) and the fact that both $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ and $\mathbb{P}_{\boldsymbol{r},\mathrm{D}}$ are discrete. The solution is given as a linear program.

The Wasserstein distance $d_{W,q}(\mathbb{P}_{\boldsymbol{r},\mathrm{B}}, \mathbb{P}_{\boldsymbol{r},\mathrm{D}})$, originally solving the *Kantorovich optimal transport problem* [18], can be interpreted as the minimal work needed to move a mass of points described via a probability distribution $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}(\boldsymbol{r})$, on the space $\mathcal{Z} \subset \mathbb{R}^p$, to a mass of points described by the probability distribution $\mathbb{P}_{\boldsymbol{r},\mathrm{D}}(\boldsymbol{r})$ on the same space, with some transportation cost ℓ. The minimization that defines $d_{W,q}$ is taken over the space of all the joint distributions Π on $\mathcal{Z} \times \mathcal{Z}$ whose marginals are $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ and $\mathbb{P}_{\boldsymbol{r},\mathrm{D}}$, respectively.

Assuming that both $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ and $\mathbb{P}_{\boldsymbol{r},\mathrm{D}}$ are discrete, we can equivalently characterize the joint distribution Π as a discrete mass *optimal transportation plan* [18]. To do this, let us consider two sets $\mathcal{N} := \{1,\dots,N\}$ and $\mathcal{T} := \{0,\dots,T-1\}$. Then, Π can be parameterized (by λ) as follows

$$\Pi_\lambda(\xi_1, \xi_2) := \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{T}} \lambda_{ij}\delta_{\{\boldsymbol{r}^{(i)}\}}(\xi_1)\delta_{\{\boldsymbol{r}(t-j)\}}(\xi_2),$$

$$\text{s.t.} \sum_{i\in\mathcal{N}}\lambda_{ij} = \frac{1}{T}, \ \forall\, j \in \mathcal{T}, \quad \sum_{j\in\mathcal{T}}\lambda_{ij} = \frac{1}{N}, \ \forall\, i \in \mathcal{N}, \tag{6}$$

$$\lambda_{ij} \geq 0, \ \forall\, i \in \mathcal{N},\ j \in \mathcal{T}. \tag{7}$$

Note that this characterizes all the joint distributions with marginals $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ and $\mathbb{P}_{\boldsymbol{r},\mathrm{D}}$, where λ is the allocation of the mass from $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ to $\mathbb{P}_{\boldsymbol{r},\mathrm{D}}$. Then, the proposed detection measure $z(t)$ in (4) reduces to the following

$$(z(t))^q := \min_{\lambda} \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{T}} \lambda_{ij}\|\boldsymbol{r}^{(i)} - \boldsymbol{r}(t-j)\|^q, \quad \text{s.t.} \ (6),\ (7), \tag{P}$$

which is a linear program over a compact polyhedral set. Therefore, the solution exists and (P) can be solved to global optimal in polynomial time by e.g., a CPLEX solver.

A. Detection and Stealthiness of Attacks

Following from the previous discussion, we now introduce an Erroneous Detection Quantification Problem, then specialize it to the Attack Detection Problem considered in this letter. In particular, we analyze the sensitivity of the proposed attack detector method for the cyber-physical system under attacks.

Problem 1 (Erroneous Detection Quantification Problem): Given the augmented system (2), the online detection procedure in Section III, and the attacker type described in Assumption 1, compute the erroneous detection probability

$$\begin{aligned} &\mathrm{Prob}(\text{erroneous detection at } t) \\ &\quad := \mathrm{Prob}(\mathrm{Alarm}(t) = 1 \mid \text{no attack})\,\mathrm{Prob}(\text{no attack}) \\ &\quad\ + \mathrm{Prob}(\mathrm{Alarm}(t) = 0 \mid \text{attack})\,\mathrm{Prob}(\text{attack}). \end{aligned}$$

Problem 1, on the computation of the erroneous detection probability, requires prior information of the attack probability Prob(attack). In this letter, we are interested in *stealthy attacks*, i.e., attacks that can avoid detection by (5). These attacks, in the worst case, can induce significant system damage before notice. We are led to the following problem.

Problem 2 (Attack Detection Problem): Given the setting of Problem 1, provide conditions that characterize stealthy attacks, i.e., attacks that contribute to $\mathrm{Prob}(\mathrm{Alarm}(t) = 0 \mid \text{attack})$, and quantify their potential impact on the system.

To remain undetected, the attacker must select $\boldsymbol{\gamma}(t)$ such that $z(t)$ is limited to within the threshold α. To quantify the effects of these attacks, let us consider an attacker sequence backward in time with length T. At time t, denote the arbitrary injected

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on October 24,2023 at 20:05:42 UTC from IEEE Xplore. Restrictions apply.

attacker sequence by $\boldsymbol{\gamma}(t-j) \in \mathbb{R}^p$, $j \in \{0, \dots, T-1\}$ (if $t-j<0$, assume $\boldsymbol{\gamma}(t-j)=0$). This sequence, together with (2), results in a detection sequence $\{\boldsymbol{r}(t-j)\}_j$ that can be used to construct detector distribution $\mathbb{P}_{\boldsymbol{r},\mathrm{D}}$ and detection measure $z(t)$. We characterize the scenarios that can occur, providing a first, partial answer to Problem 2. Then, we will come back to analyzing the impact of stealthy attacks in Section IV.

Definition 1 (Attack Detection Characterization): Assume (2) is subject to attack, i.e., $\boldsymbol{\gamma}(t) \neq \mathbf{0}$ for some $t \geq 0$.

- If $z(t) \leq \alpha$, $\forall\ t \geq 0$, then the attack is stealthy with probability one, i.e., $\mathrm{Prob}(\mathrm{Alarm}(t)=0 \mid \mathrm{attack}) = 1$.
- If $z(t) \leq \alpha$, $\forall t \leq M$, then the attack is M-step stealthy.
- If $z(t) > \alpha$, $\forall t \geq 0$, then the attack is active with probability one, i.e., $\mathrm{Prob}(\mathrm{Alarm}(t)=0 \mid \mathrm{attack}) = 0$.

For simplicity and w.l.o.g., let us assume that $\boldsymbol{\gamma}(t)$ is in form

$$\boldsymbol{\gamma}(t) := \hat{\boldsymbol{y}}^{\mathrm{o}}(t) - \boldsymbol{y}^{\mathrm{o}}(t) + \bar{\gamma}(t) = -C\boldsymbol{e}(t) - \boldsymbol{v}(t) + \bar{\gamma}(t), \quad (8)$$

where $\hat{\boldsymbol{y}}^{\mathrm{o}}(t)$, $\boldsymbol{y}^{\mathrm{o}}(t)$ are online noisy measurements of $\hat{\boldsymbol{y}}(t)$, $\boldsymbol{y}(t)$, and $\bar{\gamma}(t) \in \mathbb{R}^p$ is any vector selected by the attacker.

Lemma 2 (Stealthy Attacks Leveraging System Noise): Assume (2) is subject to attack that leverages measurements $\hat{\boldsymbol{y}}^{\mathrm{o}}(t)$ and $\boldsymbol{y}^{\mathrm{o}}(t)$ as in form (8), where $\bar{\gamma}(t)$ is stochastic and distributed as $\mathbb{P}_{\bar{\gamma}}$. If $\mathbb{P}_{\bar{\gamma}}$ is selected such that $d_{W,q}(\mathbb{P}_{\bar{\gamma}}, \mathbb{P}_{\boldsymbol{r},\mathrm{B}}) \leq \epsilon_{\mathrm{B}}$, then the attacker is stealthy with (high) probability at least $\frac{1-\Delta}{1-\beta}$, i.e., $\mathrm{Prob}(\mathrm{Alarm}(t)=0 \mid \text{ attack}) \geq \frac{1-\Delta}{1-\beta}$.[6]

Proof: Assume (2) is under attack. Leveraging the measure concentration result into the distribution $\mathbb{P}_{\bar{\gamma}}$ of the attacker,

$$\mathrm{Prob}\big(d_{W,q}(\mathbb{P}_{\bar{\gamma}}, \mathbb{P}_{\boldsymbol{r},\mathrm{D}}) \leq \epsilon_{\mathrm{D}}\big) \geq 1 - \frac{\Delta-\beta}{1-\beta},$$

which holds as $\mathbb{P}_{\boldsymbol{r},\mathrm{D}}$ is constructed using samples of $\mathbb{P}_{\bar{\gamma}}$. This together with the triangular inequality $z(t) \leq d_{W,q}(\mathbb{P}_{\boldsymbol{r},\mathrm{B}}, \mathbb{P}_{\bar{\gamma}}) + d_{W,q}(\mathbb{P}_{\boldsymbol{r},\mathrm{D}}, \mathbb{P}_{\bar{\gamma}})$, results into $z(t) \leq \alpha$ with probability at least $\frac{1-\Delta}{1-\beta}$. ■

IV. Stealthy Attack Analysis

In this section, we address the second question in Problem 2 via reachable-set analysis. In particular, note that the CPS (1) is resilient to stealthy attacks only when (1) is open-loop stable. Under this assumption, we achieve an outer-approximation of the finite-step probabilistic reachable set, quantifying the effect of the stealthy attacks and the system noise in probability.

Consider an attack sequence $\boldsymbol{\gamma}(t)$ as in (8), where $\bar{\gamma}(t) \in \mathbb{R}^p$ is any vector such that the attack remains stealthy. That is, $\bar{\gamma}(t)$ results in the detected distribution $\mathbb{P}_{\boldsymbol{r},\mathrm{D}}$, which is close to $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ as prescribed by α. This results in the representation of (2) as

$$\boldsymbol{\xi}(t+1) = \underbrace{\begin{bmatrix} A+BK & -BK \\ 0 & A \end{bmatrix}}_{H} \boldsymbol{\xi}(t) + \underbrace{\begin{bmatrix} I & 0 \\ I & -L \end{bmatrix}}_{G} \begin{bmatrix} \boldsymbol{w}(t) \\ \bar{\gamma}(t) \end{bmatrix}. \quad (9)$$

To quantify the reachable set of the system under attacks, prior information on the process noise $\boldsymbol{w}(t)$ is needed. To characterize $\boldsymbol{w}(t)$, let us assume that, similarly to the benchmark $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$, we are able to construct a noise benchmark distribution, denoted by $\mathbb{P}_{\boldsymbol{w},\mathrm{B}}$. As before,

$$\mathrm{Prob}\big(d_{W,q}(\mathbb{P}_{\boldsymbol{w},\mathrm{B}}, \mathbb{P}_{\boldsymbol{w}}) \leq \epsilon_{\boldsymbol{w},\mathrm{B}}\big) \geq 1-\beta,$$

for some $\epsilon_{\boldsymbol{w},\mathrm{B}}$. Given certain time, we are interested in where, with high probability, the state of the system can evolve from some $\boldsymbol{\xi}_0$. To do this, we consider the *M-step probabilistic reachable set* of stealthy attacks, defined as follows

$$\mathcal{R}_{\boldsymbol{x},M} := \left\{ \boldsymbol{x}(M) \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \text{system } (9) \text{ with } \boldsymbol{\xi}(0) = \boldsymbol{\xi}_0, \\ \exists\, \mathbb{P}_{\boldsymbol{w}} \ni d_{W,q}(\mathbb{P}_{\boldsymbol{w}}, \mathbb{P}_{\boldsymbol{w},\mathrm{B}}) \leq \epsilon_{\boldsymbol{w},\mathrm{B}}, \\ \exists\, \mathbb{P}_{\bar{\gamma}} \ni d_{W,q}(\mathbb{P}_{\bar{\gamma}}, \mathbb{P}_{\boldsymbol{r},\mathrm{B}}) \leq \alpha, \end{array} \right\},$$

then the true system state $\boldsymbol{x}(t)$ at time M, $\boldsymbol{x}(M)$, satisfies

$$\mathrm{Prob}\big(\boldsymbol{x}(M) \in \mathcal{R}_{\boldsymbol{x},M}\big) \geq 1-\beta,$$

where $1-\beta$ accounts for the independent restriction of the unknown distributions $\mathbb{P}_{\boldsymbol{w}}$ to be "close" to its benchmark.

The exact computation of $\mathcal{R}_{\boldsymbol{x},M}$ is intractable due to the unbounded support of the unknown distributions $\mathbb{P}_{\boldsymbol{w}}$ and $\mathbb{P}_{\bar{\gamma}}$, even if they are close to their benchmark. To ensure a tractable approximation, we follow a two-step procedure. First, we equivalently characterize the constraints on $\mathbb{P}$ by its *probabilistic support set*. Then, we outer-approximate the probabilistic support by ellipsoids, and then the reachable set by an ellipsoidal bound.

Step 1 [Probabilistic Support of $\mathbb{P}_{\bar{\gamma}} \ni d_{W,q}(\mathbb{P}_{\bar{\gamma}}, \mathbb{P}_{\boldsymbol{r},\mathrm{B}}) \leq \alpha$]: We achieve this by leveraging 1) the *Monge formulation* [18] of optimal transport, 2) the fact that $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ is discrete, and 3) results from coverage control [19], [20]. W.l.o.g., let us assume $\mathbb{P}_{\bar{\gamma}}$ is non-atomic (or continuous) and, consider the distribution $\mathbb{P}_{\bar{\gamma}}$ and $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ supported on $\mathcal{Z} \subset \mathbb{R}^p$. Let us denote by $f : \mathbb{P}_{\bar{\gamma}} \mapsto \mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ the *transport map* that assigns mass over $\mathcal{Z}$ from $\mathbb{P}_{\bar{\gamma}}$ to $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$. The Monge formulation aims to find an optimal transport map that minimizes the transportation cost ℓ as follows

$$d_{M,q}(\mathbb{P}_{\bar{\gamma}}, \mathbb{P}_{\boldsymbol{r},\mathrm{B}}) := \left(\min_{f} \int_{\mathcal{Z}} \ell^q(\xi, f(\xi)) \mathbb{P}_{\bar{\gamma}}(\xi) d\xi \right)^{1/q}.$$

It is known that if an optimal transport map $f^\star$ exists, then the optimal transportation plan $\Pi^\star$ exists and the two problems $d_{M,q}$ and $d_{W,q}$ coincide.[7] In our setting, for strongly convex ℓ^p, and by the fact that $\mathbb{P}_{\bar{\gamma}}$ is absolutely continuous, a unique optimal transport map can indeed be guaranteed,[8] and therefore, $d_{M,q} = d_{W,q}$. Let us now consider any transport map f of $d_{M,q}$, and define a partition of the support of $\mathbb{P}_{\bar{\gamma}}$ by

$$W_i := \{\boldsymbol{r} \in \mathcal{Z} \mid f(\boldsymbol{r}) = \boldsymbol{r}^{(i)}\},\ i \in \mathcal{N},$$

where $\boldsymbol{r}^{(i)}$ are samples in Ξ_{B}, which generate $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$. Then, we equivalently rewrite the objective function defined in $d_{M,q}$, as

$$\mathcal{H}(\mathbb{P}_{\bar{\gamma}}, W_1, \dots, W_N) := \sum_{i=1}^{N} \int_{W_i} \ell^q(\xi, \boldsymbol{r}^{(i)}) \mathbb{P}_{\bar{\gamma}}(\xi) d\xi,$$
$$\text{s.t.} \quad \int_{W_i} \mathbb{P}_{\bar{\gamma}}(\xi) d\xi = \frac{1}{N},\ \forall\, i \in \mathcal{N}, \quad (10)$$

[6] Note that $\alpha > \epsilon_{\mathrm{B}}$, which allows the attacker to select $\mathbb{P}_{\bar{\gamma}}$ with $\epsilon_{\mathrm{B}} < d_{W,q}(\mathbb{P}_{\bar{\gamma}}, \mathbb{P}_{\boldsymbol{r},\mathrm{B}}) \leq \alpha$. However, the probability of being stealthy can be indefinitely low, with the range $[0, \frac{1-\Delta}{1-\beta}]$.

[7] This is because the Kantorovich transport problem is the tightest relaxation of the Monge transport problem. See [18] for details.

[8] The Monge formulation is not always well-posed, i.e., there exists optimal transportation plans $\Pi^\star$ while transport map does not exist [18].

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on October 24,2023 at 20:05:42 UTC from IEEE Xplore. Restrictions apply.

where the i^{th} constraints come from the fact that a transport map f should lead to consistent calculation of set volumes under $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ and $\mathbb{P}_{\bar{\gamma}}$, respectively. This results in the following equivalent problem to $d_{M,q}$ as

$$(d_{M,q}(\mathbb{P}_{\bar{\gamma}}, \mathbb{P}_{\boldsymbol{r},\mathrm{B}}))^q := \min_{W_i, i \in \mathcal{N}} \ \mathcal{H}(\mathbb{P}_{\bar{\gamma}}, W_1, \ldots, W_N), \quad \text{s.t. } (10). \tag{P1}$$

Given the distribution $\mathbb{P}_{\bar{\gamma}}$, Problem (P1) reduces to a load-balancing problem as in [20]. Let us define the Generalized Voronoi Partition (GVP) of $\mathcal{Z}$ associated to the sample set Ξ_{B} and weight $\omega \in \mathbb{R}^N$, for all $i \in \mathcal{N}$, as

$$\begin{aligned}&\mathcal{V}_i(\Xi_{\mathrm{B}}, \omega)\\ &:= \{\xi \in \mathcal{Z} \mid \|\xi - \boldsymbol{r}^{(i)}\|^q - \omega_i \le \|\xi - \boldsymbol{r}^{(j)}\|^q - \omega_j, \ \forall j \in \mathcal{N}\}.\end{aligned}$$

It has been established that the optimal Partition of (P1) is the GVP [20, Proposition V.1]. Further, the standard Voronoi partition, i.e., the GVP with equal weights $\bar{\omega} := \mathbf{0}$, results in a lower bound of (P1), when constraints (10) are removed [19], and therefore that of $d_{M,q}$. We denote this lower bound as $L(\mathbb{P}_{\bar{\gamma}}, \mathcal{V}(\Xi_{\mathrm{B}}))$, and use this to quantify a probabilistic support of $\mathbb{P}_{\bar{\gamma}}$. Let us consider the support set

$$\Omega(\Xi_{\mathrm{B}}, \epsilon) := \cup_{i \in \mathcal{N}} \Big(\mathcal{V}_i(\Xi_{\mathrm{B}}) \cap \mathbb{B}_{\epsilon}(\boldsymbol{r}^{(i)}) \Big),$$

where $\mathbb{B}_{\epsilon}(\boldsymbol{r}^{(i)}) := \{\boldsymbol{r} \in \mathbb{R}^p \mid \|\boldsymbol{r} - \boldsymbol{r}^{(i)}\| \le \epsilon\}$.

Lemma 3 (Probabilistic Support): Let $\epsilon > 0$ and let $\mathbb{P}_{\bar{\gamma}}$ be a distribution such that $L(\mathbb{P}_{\bar{\gamma}}, \mathcal{V}(\Xi_{\mathrm{B}})) \le \epsilon^q$. Then, for any given $s > 1$, at least $1 - 1/s^q$ portion of the mass of $\mathbb{P}_{\bar{\gamma}}$ is supported on $\Omega(\Xi_{\mathrm{B}}, s\epsilon)$, i.e., $\int_{\Omega(\Xi_{\mathrm{B}}, s\epsilon)} \mathbb{P}_{\bar{\gamma}}(\xi) d\xi \ge 1 - 1/s^q$.

Proof: Suppose otherwise, i.e., $\int_{\mathbb{R}^p \setminus \Omega(\Xi_{\mathrm{B}}, s\epsilon)} \mathbb{P}_{\bar{\gamma}}(\xi) d\xi = 1 - \int_{\Omega(\Xi_{\mathrm{B}}, s\epsilon)} \mathbb{P}_{\bar{\gamma}}(\xi) d\xi > 1/s^q$. Then,

$$\begin{aligned} L(\mathbb{P}_{\bar{\gamma}}, \mathcal{V}(\Xi_{\mathrm{B}})) &\ge \int_{\mathbb{R}^p \setminus \Omega(\Xi_{\mathrm{B}}, s\epsilon)} \|\xi - \boldsymbol{r}^{(i)}\|^q \mathbb{P}_{\bar{\gamma}}(\xi) d\xi, \\ &\ge s^q \epsilon^q \int_{\mathbb{R}^p \setminus \Omega(\Xi_{\mathrm{B}}, s\epsilon)} \mathbb{P}_{\bar{\gamma}}(\xi) d\xi > \epsilon^q, \text{ contradiction.} \end{aligned}$$

■

In this way, the support $\Omega(\Xi_{\mathrm{B}}, 2\alpha)$ contains at least $1 - 1/2^q$ of the mass of all the distributions $\mathbb{P}_{\bar{\gamma}}$ such that $d_{W,q}(\mathbb{P}_{\bar{\gamma}}, \mathbb{P}_{\boldsymbol{r},\mathrm{B}}) \le \alpha$. Equivalently, we have $\mathrm{Prob}(\bar{\gamma} \in \Omega(\Xi_{\mathrm{B}}, 2\alpha)) \ge 1 - 1/2^q$, where the random variable $\bar{\gamma}$ has a distribution $\mathbb{P}_{\bar{\gamma}}$ such that $d_{W,q}(\mathbb{P}_{\bar{\gamma}}, \mathbb{P}_{\boldsymbol{r},\mathrm{B}}) \le \alpha$. We characterize $\mathbb{P}_{\boldsymbol{w}}$ similarly.

Step 2 (Outer-approximation of $\mathcal{R}_{\boldsymbol{x},M}$): Making use of the probabilistic support, we can now obtain a finite-dimensional characterization of the probabilistic reachable set, as follows

$$\mathcal{R}_{\boldsymbol{x},M} := \left\{ \boldsymbol{x}(M) \in \mathbb{R}^n \;\middle|\; \begin{array}{l} \text{system } (9),\ \boldsymbol{\xi}(0) = \boldsymbol{\xi}_0, \\ \boldsymbol{w} \in \Omega(\Xi_{\boldsymbol{w},\mathrm{B}}, 2\epsilon_{\boldsymbol{w},\mathrm{B}}), \\ \bar{\gamma} \in \Omega(\Xi_{\mathrm{B}}, 2\alpha) \end{array} \right\},$$

and the true system state $\boldsymbol{x}(t)$ at time M, $\boldsymbol{x}(M)$, satisfies $\mathrm{Prob}(\boldsymbol{x}(M) \in \mathcal{R}_{\boldsymbol{x},M}) \ge (1-\beta)(1 - 1/2^q)^2$. Note that, if the CPS (1) is open-loop unstable, so does (9). This leads to vulnerable CPS to stealthy sensor attacks. That is, almost surely any stealthy attack $\boldsymbol{\gamma}(t)$ in form (8) inflicts significant damage of the system with an unbounded reachable set, i.e., $\exists\ \boldsymbol{x}(M) \in \mathcal{R}_{\boldsymbol{x},M}$ such that $\boldsymbol{x}(M) \to \infty$ as $M \to \infty$. Many works focus on the tractable evolution of geometric shapes when (1) is stable and resilient to stealthy attacks,[9] e.g., [9], [13]. Here, we follow [13] and propose outer ellipsoidal bounds for $\mathcal{R}_{\boldsymbol{x},M}$. Let $Q_{\boldsymbol{w}}$ be the positive-definite shape matrix such that $\Omega(\Xi_{\boldsymbol{w},\mathrm{B}}, \epsilon_{\boldsymbol{w},\mathrm{B}}) \subset \mathcal{E}_{\boldsymbol{w}} := \{\boldsymbol{w} \mid \boldsymbol{w}^\top Q_{\boldsymbol{w}} \boldsymbol{w} \le 1\}$. Similarly, we denote $Q_{\bar{\gamma}}$ and $\mathcal{E}_{\bar{\gamma}}$ for that of $\bar{\gamma}$. We now state the lemma, that applies [13, Proposition 1] for our case.

Lemma 4 (Outer bounds of $\mathcal{R}_{\boldsymbol{x},M}$): Given any $a_0 \in (0, 1)$, we claim $\mathcal{R}_{\boldsymbol{x},M} \subset \mathcal{E}(Q) := \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{\xi}^\top Q \boldsymbol{\xi} \le a_0^M \boldsymbol{\xi}_0{}^\top Q \boldsymbol{\xi}_0 + \frac{(2-a_0)(1-a_0^M)}{1-a_0}\}$, with Q satisfying

$$Q \succ 0, \quad \begin{bmatrix} a_0 Q & H^\top Q & \mathbf{0} \\ QH & Q & QG \\ \mathbf{0} & G^\top Q & W \end{bmatrix} \succeq 0, \tag{11}$$

where H, G are that in (9) and

$$W = \begin{bmatrix} (1 - a_1) Q_{\boldsymbol{w}} & \mathbf{0} \\ \mathbf{0} & (1 - a_2) Q_{\bar{\gamma}} \end{bmatrix}, \quad \text{for some } a_1 + a_2 \ge a_0,\ a_1, a_2 \in (0, 1). \tag{12}$$

A tight reachable set bound can be now derived by solving

$$\min_{Q, a_1, a_2} \ -\log\det(Q), \quad \text{s.t. } (11),\ (12), \tag{P2}$$

which is a convex semidefinite program, solvable via e.g., SeDuMi [21]. Note that the probabilistic reachable set is

$$\mathcal{R}_{\boldsymbol{x}} := \cup_{M=1}^{\infty} \mathcal{R}_{\boldsymbol{x},M},$$

which again can be approximated via $Q^\star$ solving (P2) for[10]

$$\mathcal{R}_{\boldsymbol{x}} \subset \mathcal{E}(Q^\star) = \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{\xi}^\top Q^\star \boldsymbol{\xi} \le \frac{(2 - a_0)}{1 - a_0}\}.$$

V. Simulations

In this section, we demonstrate the performance of the proposed attack detector, illustrating its distributional robustness w.r.t. the system noise. Then, we consider stealthy attacks as in (8) and analyze their impact by quantifying the probabilistic reachable set and outer-approximation bound.

Consider the stochastic system (2), given as

$$\begin{aligned} &A = \begin{bmatrix} 1.00 & 0.10 \\ -0.20 & 0.75 \end{bmatrix}, \ B = \begin{bmatrix} 0.10 \\ 0.20 \end{bmatrix}, \ L = \begin{bmatrix} 0.23 \\ -0.20 \end{bmatrix}, \\ &C = \begin{bmatrix} 1 & 0 \end{bmatrix}, \ K = \begin{bmatrix} -0.13 & 0.01 \end{bmatrix}, \ n = 2, \ m = p = 1, \\ &\boldsymbol{w}_1 \sim \mathcal{N}(-0.25, 0.02) + \mathcal{U}(0, 0.5), \ \boldsymbol{v} \sim \mathcal{U}(-0.3, 0.3), \\ &\boldsymbol{w}_2 \sim \mathcal{N}(0, 0.04) + \mathcal{U}(-0.2, 0.2), \end{aligned}$$

where $\mathcal{N}$ and $\mathcal{U}$ represent the normal and uniform distributions, respectively. We consider $N = 10^3$ benchmark samples for $\mathbb{P}_{\boldsymbol{r},\mathrm{B}}$ and $T = 10^2$ real-time samples for $\mathbb{P}_{\boldsymbol{r},\mathrm{D}}$. We select $q = 1$, $\beta = 0.01$ and false alarm rate $\Delta = 0.05$. We select the prior information of the system noise via $a = 1.5$, $c_1 = 1.84 \times 10^6$ and $c_2 = 12.5$. Using the measure-of-concentration results, we determine the detector threshold to be $\alpha = 0.158$. In the normal system operation (no attack),

[9] If (1) is unstable, we either need extra protected sensors or benchmark data of the state estimate $\hat{\boldsymbol{x}}$, $\mathbb{P}_{\hat{\boldsymbol{x}},\mathrm{B}}$, to ensure effective stealthy attack detection.

[10] The set $\mathcal{R}_{\boldsymbol{x}}$ is in fact contained in the projection of $\mathcal{E}(Q^\star)$ onto the state subspace, i.e., $\mathcal{R}_{\boldsymbol{x}} \subset \{\boldsymbol{x} \mid \boldsymbol{x}^\top (Q_{xx} - Q_{xe} Q_{ee}^{-1} Q_{xe}{}^\top) \boldsymbol{x} \le \frac{(2-a_0)}{1-a_0}\}$ with $Q^\star := \begin{bmatrix} Q_{xx} & Q_{xe} \\ Q_{xe}{}^\top & Q_{ee} \end{bmatrix}$. See [13] for details.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on October 24,2023 at 20:05:42 UTC from IEEE Xplore. Restrictions apply.

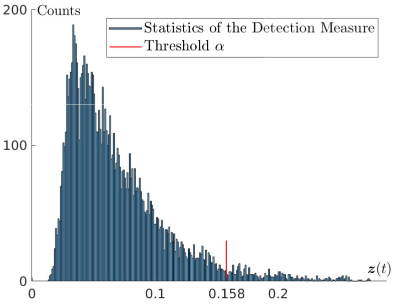

Fig. 1. Statistics of z.

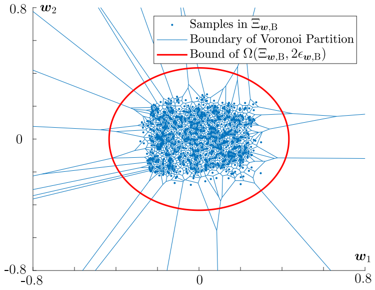

Fig. 2. Probabilistic Support of $\mathbb{P}_{w}$.

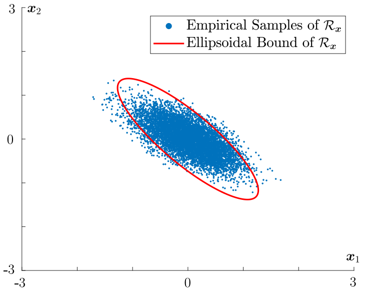

Fig. 3. Empirical and Bound of $\mathcal{R}_{x}$.

we run the online detection procedure for 10^4 time steps and draw the distribution of the computed detection measure $z(t)$ as in Fig. 1. We verify that the false alarm rate is 3.68%, within the required rate $\Delta = 5\%$. When the system is subject to stealthy attacks, we assume $\boldsymbol{\xi}_0 = \mathbf{0}$ and visualize the Voronoi partition $\mathcal{V}(\Xi_{w,\mathrm{B}})$ (convex sets with blue boundaries) of the probabilistic support $\Omega(\Xi_{w,\mathrm{B}}, \epsilon_{w,\mathrm{B}})$ and its estimated ellipsoidal bound (red line) as in Fig. 2. Further, we demonstrate the impact of the stealthy attacks (8) with $a_0 = 0.85$, as in Fig. 3. We used 10^4 empirical points of $\mathcal{R}_{x}$ as its estimate and provided an ellipsoidal bound of $\mathcal{R}_{x}$ computed by solution of (P2). It can be seen that the proposed probabilistic reachable set effectively captures the reachable set in probability. Due to the space limits, we omit the comparison of our approach to the existing ones, such as the classical χ^2 detector in [9] and the CUMSUM procedure [8]. However, the difference should be clear: our proposed approach is robust w.r.t. noise distributions while others leverage the moment information up to the second order, which only capture sufficient information about certain noise distributions, e.g., Gaussian.

VI. Conclusion

A novel detection measure was proposed to enable distributionally robust detection of attacks w.r.t. unknown, and light-tailed system noise. The proposed detection measure restricts the behavior of the stealthy attacks, whose impact was quantified via reachable-set analysis.

References

[1] A. Cardenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, and S. Sastry, "Challenges for securing cyber physical systems," in *Proc. Workshop Future Directions Cyber Phys. Syst.*, vol. 5, 2009, pp. 1–6.

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

[3] S. Amin, A. Cardenas, and S. S. Sastry, "Safe and secure networked control systems under denial-of-service attacks," in *Hybrid Systems: Computation and Control*. Heidelberg, Germany: Springer, 2009, pp. 31–45.

[4] F. Miao, Q. Zhu, M. Pajic, and G. J. Pappas, "Coding schemes for securing cyber-physical systems against stealthy data injection attacks," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 1, pp. 106–117, Mar. 2017.

[5] C. Bai, F. Pasqualetti, and V. Gupta, "Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs," *Automatica*, vol. 82, pp. 251–260, Aug. 2017.

[6] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Proc. 47th Annu. Allerton Conf. Commun. Control Comput.*, Monticello, IL, USA, Sep. 2009, pp. 911–918.

[7] M. Zhu and S. Martínez, "On the performance analysis of resilient networked control systems under replay attacks," *IEEE Trans. Autom. Control*, vol. 59, no. 3, pp. 804–808, Mar. 2014.

[8] J. Milošević, D. Umsonst, H. Sandberg, and K. Johansson, "Quantifying the impact of cyber-attack strategies for control systems equipped with an anomaly detector," in *Proc. Eur. Control Conf.*, Limassol, Cyprus, 2018, pp. 331–337.

[9] Y. Mo and B. Sinopoli, "On the performance degradation of cyber-physical systems under stealthy integrity attacks," *IEEE Trans. Autom. Control*, vol. 61, no. 9, pp. 2618–2624, Sep. 2016.

[10] S. Mishra, Y. Shoukry, N. Karamchandani, S. Diggavi, and P. Tabuada, "Secure state estimation against sensor attacks in the presence of noise," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 1, pp. 49–59, Mar. 2017.

[11] C. Murguia, N. V. de Wouw, and J. Ruths, "Reachable sets of hidden CPS sensor attacks: Analysis and synthesis tools," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 2088–2094, 2017.

[12] C. Bai, F. Pasqualetti, and V. Gupta, "Security in stochastic control systems: Fundamental limitations and performance bounds," in *Proc. Amer. Control Conf.*, Chicago, IL, USA, 2015, pp. 195–200.

[13] C. Murguia, I. Shames, J. Ruths, and D. Nesic, "Security metrics of networked control systems under sensor attacks," 2018. [Online]. Available: arXiv:1809.01808.

[14] V. Renganathan, N. Hashemi, J. Ruths, and T. Summers, "Distributionally robust tuning of anomaly detectors in cyber-physical systems with stealthy attacks," 2019. [Online]. Available: arXiv:1909.12506.

[15] D. Li and S. Martínez, "High-confidence attack detection via Wasserstein-metric computations," 2020. [Online]. Available: arXiv:2003.07880.

[16] A. R. Teel, J. P. Hespanha, and A. Subbaraman, "Equivalent characterizations of input-to-state stability for stochastic discrete-time systems," *IEEE Trans. Autom. Control*, vol. 59, no. 2, pp. 516–522, Feb. 2014.

[17] N. Fournier and A. Guillin, "On the rate of convergence in Wasserstein distance of the empirical measure," *Probab. Theory Related Fields*, vol. 162, nos. 3–4, pp. 707–738, 2015.

[18] F. Santambrogio, *Optimal Transport for Applied Mathematicians*. Cham, Switzerland: Springer, 2015.

[19] F. Bullo, J. Cortés, and S. Martínez, *Distributed Control of Robotic Networks* (Applied Mathematics Series). Princeton, NJ, USA: Princeton Univ. Press, 2009.

[20] J. Cortés, "Coverage optimization and spatial load balancing by robotic sensor networks," *IEEE Trans. Autom. Control*, vol. 55, no. 3, pp. 749–754, Mar. 2010.

[21] J. Sturm, "Using SeDumi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optim. Methods Softw.*, vol. 11, nos. 1–4, pp. 625–653, 1999.

Authorized licensed use limited to: Univ of Calif San Diego. Downloaded on October 24,2023 at 20:05:42 UTC from IEEE Xplore. Restrictions apply.