Safety of Linear Systems under Severe Sensor Attacks

Xiao Tan, Pio Ong, Paulo Tabuada, and Aaron D. Ames

Abstract— **Cyber-physical systems can be subject to sensor attacks, e.g., sensor spoofing, leading to unsafe behaviors. This paper addresses this problem in the context of linear systems when an omniscient attacker can spoof several system sensors at will. In this adversarial environment, existing results have derived necessary and sufficient conditions under which the state estimation problem has a unique solution. In this work, we consider a severe attacking scenario when such conditions do not hold. To deal with potential state estimation uncertainty, we derive an exact characterization of the set of all possible state estimates. Using the framework of control barrier functions, we propose design principles for system safety in offline and online phases. For the offline phase, we derive conditions on safe sets for all possible sensor attacks that may be encountered during system deployment. For the online phase, with past system measurements collected, a quadratic program-based safety filter is proposed to enforce system safety. A 2D-vehicle example is used to illustrate the theoretical results.**

I. Introduction

Cyber-physical systems (CPS), as an integration of computation, communication, and physical processes, range from small-scale applications, such as cars, to large-scale infrastructure, such as smart grids and water distribution systems. Because of their close interaction of physical and computational components, CPS are prone to attacks in both cyber and physical domains. Prominent examples of such attacks include the Stuxnet malware [1] targeting on a process control system. Previous works also demonstrate that attacks may come from the physical domain [2], [3]. In [2], the authors spoof the velocity measurements of a vehicle, which, when intervened by anti-lock braking systems, may cause the drivers to lose control of their vehicles. Motivated by these real-world examples, researchers have explored various attacking and defense strategies for CPS, including, e.g., denial-of-service [4], replay attacks [5], man-in-the-middle [6], and false data injection [7].

In this paper, we consider a scenario similar to those in [2], [3] wherein certain measurements of the CPS are compromised by an attacker. In the setting of linear systems, we adopt a general attack model that imposes no limitations on the magnitude, statistical properties, or temporal evolution requirements on the attack signal. Rather, we only assume an upper bound on the number of attacked sensors. Most existing results in this setting have focused on recovering the system state from compromised measurement data, known as the secure state reconstruction problem. For discrete-time linear systems, [8] derives several necessary and sufficient conditions for the uniqueness of solutions to this problem, which are further refined in [9]. There, the condition is posed as a sparse observability property of the CPS. An equivalent condition for continuous-time linear systems is given in [10]. Recently, [11] shows that finding this unique solution is NP-hard in general.

An important question naturally arises for CPS in this adversarial scenario: can we ensure *safety* of the system, and thereby avoid catastrophic results through active control, even when certain sensors are compromised? Safety in control systems usually refers to the property that system trajectories can be made to stay within a safe set via feedback. Compromised sensor measurements will negatively affect or mislead the state estimates, and thus complicate safe control design. Motivated by these challenges, [12] and [13] have explored safe control designs in this setting.In [12], the sensor attack signals are taken to be bounded, and a set of safeguarded sensors are assumed available for the design of the so-called secondary controller. In [13], a finite attack pattern is assumed, and each pattern corresponds to a particular subset of compromised sensors. A fault identification scheme is proposed by implementing a large number of extended Kalman filters simultaneously. They further make assumptions on the sensor attacks so that the state estimate error is bounded in probability. The assumptions made by these works may not hold true under the aforementioned sensor attack model that we adopt.

In recent years, control barrier functions (CBF) [14] have gained popularity as a framework for safety-critical control—providing Lyapunov-like necessary and sufficient conditions for forward set invariance. A strength of this approach is its ability to incorporate (bounded) uncertainty, including uncertainty in the input [15], uncertainty in the state [16], and measurement [17]. There have also been non-deterministic characterizations of safety in the context of risk-adverse CBFs [18], [19], along with stochastic CBFs [20], where the possible state/perturbation follows a certain stochastic distribution. Yet when unbounded, non-stochastic, intelligent sensor attacks are performed by an omniscient attacker, the state estimation error does not satisfy the assumptions of these works. Note that CBFs have been applied to address other security issues, such as privacy preservation [21] and safety in the presence of faulty sensors [13].

In this work, we focus on safety guarantees for CPS subject to general sensor attacks described above. In particular, we consider scenarios where the solution to the secure state

This work is supported by TII under project #A6847.

Xiao Tan, Pio Ong, and Aaron D. Ames are with the the Department of Mechanical and Civil Engineering, California Institute of Technology, Pasadena, CA 91125, USA (Email: xiaotan, pioong, ames@caltech.edu).

Paulo Tabuada is with the Department of Electrical and Computer Engineering at University of California, Los Angeles, CA 90095, USA (Email: tabuada@ucla.edu).

reconstruction problem may not be unique. Our contributions are summarized as follow:

1) We provide an exact characterization of the set of possible solutions to the secure state reconstruction problem for linear discrete-time systems.
2) We outline design principles for safe sets in the offline phase for the worst-case attacking scenario under a mild sparse observability assumption.
3) We propose an online safe control scheme that provides safety guarantees in the presence of possibly unbounded state estimation error.

Ultimately, this paper presents a general characterization of the safety of linear systems subject to *severe sensor attacks*.

Notation: For $w \in \mathbb{N}$, define $[w] := \{1, 2, ..., w\}$. The cardinality of a set $\mathcal{I}$ is denoted by $|\mathcal{I}|$. Given a $w \in \mathbb{N}$, a k-combination from $[w]$ is a subset of $[w]$ with cardinality k. Denote by $\mathbb{C}_w^k$ the set of all k-combinations from $[w]$. For a matrix $C \in \mathbb{R}^{w\times n}$ and an index set $\Gamma \subseteq [w]$, denote by C_Γ the matrix obtained from C by removing all the rows with indices not in Γ. For a point $x \in \mathbb{R}^n$, a set $\mathcal{X} \subseteq \mathbb{R}^n$, and a matrix $A \in \mathbb{R}^{n\times n}$, we define $\|x\|_{\mathcal{X}} := \min_v \|x - v\|$ s. t. $v \in \mathcal{X}$, and $A(\mathcal{X}) := \{y \in \mathbb{R}^n : y = Ax, x \in \mathcal{X}\}$. Minkowski summation $\mathcal{X}_1 + \mathcal{X}_2$ for two sets $\mathcal{X}_1, \mathcal{X}_2$ is defined as $\{x_1 + x_2, x_1 \in \mathcal{X}_1, x_2 \in \mathcal{X}_2\}$. When no confusion arises, with slight abuse of notation, we use interchangeably a vector $x \in \mathbb{R}^n$ and the singleton set $\{x\}$.

II. Problem formulation

Consider a discrete-time linear system under sensor attacks

$$\begin{aligned} \text{(dynamics)} \quad & x(\tau+1) = Ax(\tau) + Bu(\tau), \\ \text{(measurement)} \quad & y(\tau) = Cx(\tau) + e(\tau), \\ \text{(safe set)} \quad & \mathcal{C} = \{x \in \mathbb{R}^n : h(x) := Hx + q \geq 0\}, \end{aligned} \tag{1}$$

where $x(\tau) \in \mathbb{R}^n$, $u(\tau) \in \mathbb{R}^m$, $y(\tau) \in \mathbb{R}^p$, $e(\tau) \in \mathbb{R}^p$, represent the system state, the control input, the output measurement, and the attacking signal on the sensors, respectively. Regarding the attacks, $e_i(\tau)$ is nonzero whenever a sensor $i \in [p]$ is under attack at time τ. A safe set $\mathcal{C}$ is the set of states that are permitted along the system trajectories. Here, we specify the set with a vector-valued linear function $h : \mathbb{R}^n \to \mathbb{R}^l$, representing a polytopic safe set. For notation simplicity, we refer to the input-output data $(u(0), u(1), \ldots, u(t-1), y(0), y(1), \ldots, y(t)) \in \mathbb{R}^{m(t-1)+pt}$ collected after receiving the latest measurement and before choosing a control input at time t as $\mathcal{D}_t$.

We adopt the sensor attack model presented in [8], [9], and assume the following throughout the paper.

Assumption 1. The attacker has the full knowledge of the system, including the states, the dynamics (1), and our defense strategy. It may choose s out of the p sensors to attack, the choice of which remains unchanged for the duration considered. For these sensors, the attacker can adjust the corresponding $e_i(\tau)$ to any value.

We now introduce the notions of system safety and control barrier functions. Note that we restrict the class of barrier functions to be linear.

Definition 1 (System safety)**.** A system can be rendered safe from time t if there exists an input sequence $\{u(\tau)\}_{\tau\geq t}$ such that the state $x(\tau)$ remains in the safe set $\mathcal{C}$ for all $\tau \geq t$.

Definition 2 (Control barrier function [22])**.** The function $h : x \mapsto Hx + q$ with $H \in \mathbb{R}^{l\times n}$ and $q \in \mathbb{R}^l$ is called a *(discrete-time) control barrier function* for system (1) if, for some $\gamma \in (0, 1)$, for each $x \in \mathbb{R}^n$, there exists an input $u \in \mathbb{R}^m$ such that

$$H(Ax + Bu) + q \geq (1 - \gamma)(Hx + q). \tag{2}$$

It has been shown in [14], [22] that if $u(\tau)$ is chosen to satisfy the condition in (2) along trajectories, then the safe set $\mathcal{C}$ is forward invariant and asymptotically stable. Hereafter we refer to (2) as the CBF condition.

When system sensors are under attack, one immediate question is "can a good state estimate be retained?" Related to this are the following notions.

Definition 3 (r-sparse observability)**.** System (1) is *r-sparse observable* if the pair (A, C_Γ) is observable for any index set $\Gamma \in \mathbb{C}_p^{p-r}$.

Definition 4 (Plausible initial states)**.** Given input-output data $\mathcal{D}_t$, we call x_0 *a plausible initial state* if there exists $e(\tau), \tau = 0, 1, ..., t$, satisfying Assumption 1 such that the first two equations in (1) hold with $x(0) = x_0$. We denote the set of all plausible initial states at time t by $\mathcal{X}_t^0$.

It is known that there exists a unique solution to the secure state reconstruction problem, i.e., the set of all plausible initial state $\mathcal{X}_t^0$ is a singleton if and only if the system is $2s$-sparse observable [9]. In this work, we investigate the case beyond the $2s$-sparse observability condition, which is referred to as the *severe sensor attacks case*. In this case, the plausible initial states may not be unique. As will be shown later, this ambiguity complicates system safety analysis.

In this work, we consider the system safety guarantee under severe sensor attacks both in offline and online phases. For the offline phase, we characterize the set of plausible states under the worst-case sensor attack, and identify sets in the state space for which the system can be rendered safe. For the online phase, where measurements are collected for the first t steps, we derive control barrier function-based conditions that guarantee system safety from time t. This is formally stated as follows.

Problem 1 (Worst-case sensor attack). Given system matrices A, B, C, and the number of sensor attacks s, derive conditions on H and q such that the system can be rendered safe under all possible sensor attacks.

Problem 2 (Fixed yet unknown sensor attack). Given system matrices A, B, C, the number of sensor attacks s, and the input-output data $\mathcal{D}_t$, $t \geq n$, derive conditions on H, q, and the input sequence $\{u(\tau)\}_{\tau\geq t}$ such that the system is safe.

III. Secure state estimation and safe control

In this section, we present our main results on system safety under sensor attacks. We approach this problem in two steps: 1) secure state estimation and 2) safe control

design. First, we present results on the characterization of the set of all plausible states. We then derive the conditions that guarantee that all state trajectories emanating from the set of all plausible initial states remain within the safe set. We highlight that the set of plausible initial states $\mathcal{X}_t^0$ may contain multiple elements.

A. Characterization of plausible states

In this section, we only consider the case $t \geq n-1$. When $t < n-1$, some states may become plausible states due to the lack of measurements, which is not the focus of this work.

Following the derivations in [9], by stacking the measurement history per sensor, we can rewrite the equations in (1) in a compact form as follows:

$$Y_i = \mathcal{O}_i x_0 + E_i, \ \ i = 1, 2, ..., p, \tag{3}$$

where the matrices are defined in (5) at the bottom of this page, and $Y_i = \widetilde{Y}_i - F_i U$. We establish the following result on the set $\mathcal{X}_t^0$.

Proposition 1. *Given system* (1)*, the maximal number of attacked sensors s, and the input-output data $\mathcal{D}_t$, we have:*

$$\mathcal{X}_t^0 = \bigcup_{\forall \ \{i_1,...,i_{p-s}\} \in \mathbb{C}_p^{p-s}} \left(\mathcal{X}_t^{0,i_1} \cap \mathcal{X}_t^{0,i_2} \cap ... \cap \mathcal{X}_t^{0,i_{p-s}} \right) \tag{4}$$

where $\mathcal{X}_t^{0,i} = \{x : \mathcal{O}_i x = Y_i\}$ for $i \in [p]$.

Proof. First we show every element in $\mathcal{X}_t^0$ is a plausible initial state under s sensor attacks. Consider any $x_0 \in \mathcal{X}_t^0$. The state x_0 must belong in one of the sets in the union in (4). That is, there exists a set of $p-s$ indices $\Gamma = \{i_1, ..., i_{p-s}\} \in \mathbb{C}_p^{p-s}$ such that $\mathcal{O}_i x_0 = Y_i$ for $i \in \Gamma$. Then, because $|\Gamma| = p-s$, there are precisely s sensor indices left, $[p] \setminus \Gamma$. We can choose $E_j = Y_j - \mathcal{O}_j x, \ \forall j \in [p] \setminus \Gamma$ so that (3) holds for each index j. Thus, there exists an attacking signal for which $E_i = 0$ for $p-s$ indices and nonzero for s indices, thereby satisfying Assumption 1, such that (3) holds, and hence x_0 is a plausible initial state.

Now we show that no state other than those in $\mathcal{X}_t^0$ in (4) is a plausible initial state with the measurements as per Definition 4. Suppose there exists a plausible state $x_0 \notin \mathcal{X}_t^0$. Then, there must exist a set of indices $\Gamma \in \mathbb{C}_p^{p-s}$ (the set of intact sensors) such that $E_i = 0$. As a result, we have $\mathcal{O}_i x_0 = Y_i, i \in \Gamma$ from (3). This, however, suggests that x_0 belongs to one of the sets in the union in (4), so it belongs to the set $\mathcal{X}_t^0$, which is a contradiction. □

Following (4), one observes that, in general, the set of plausible initial states $\mathcal{X}_t^0$ is a union of affine subspaces. We use this result to derive the set of plausible states at time t by forward propagating $\mathcal{X}_t^0$ through the system model (1). To ease the notation, for a combination $\Gamma = \{i_1, i_2, ..., i_{p-s}\} \in \mathbb{C}_p^{p-s}$, let $\mathcal{X}_t^{0,\Gamma} = \{x : \mathcal{O}_\Gamma x = Y_\Gamma\}$, where

$$\mathcal{O}_\Gamma := \begin{bmatrix} \mathcal{O}_{i_1} \\ \mathcal{O}_{i_2} \\ \vdots \\ \mathcal{O}_{i_{p-s}} \end{bmatrix}, \ Y_\Gamma := \begin{bmatrix} Y_{i_1} \\ Y_{i_2} \\ \vdots \\ Y_{i_{p-s}} \end{bmatrix}. \tag{6}$$

Furthermore, we use the notation $\mathcal{X}_t^\tau$ and $\mathcal{X}_t^{\tau,\Gamma}$ for the set of all plausible states at time τ, starting from $\mathcal{X}_t^0$ and $\mathcal{X}_t^{0,\Gamma}$, respectively. We will omit the subscript t when it is clear the measurements considered are from the first t steps. We now characterize the set $\mathcal{X}_t^t$.

Corollary 1. *Under the premises of Proposition 1, we have:*

$$\mathcal{X}^t = \bigcup_{\forall \Gamma \in \mathbb{C}_p^{p-s}} \mathcal{X}^{t,\Gamma} \tag{7}$$

where, for each $\Gamma \in \mathbb{C}_p^{p-s}$,

$$\mathcal{X}^{t,\Gamma} = A^t(\mathcal{X}^{0,\Gamma}) + A^{t-1}Bu(0) + ... + Bu(t-1). \tag{8}$$

Proof. The result follows from a direct calculation of the system dynamics $\mathcal{X}_t^{\tau+1,\Gamma} = A(\mathcal{X}_t^{\tau,\Gamma}) + Bu(\tau)$ for all $\tau \in \{0, 1, ..., t-1\}$. □

B. Offline phase safe set design

In this subsection, we establish conditions on (A, B, C, s, H, q) that provide safety guarantees under all possible attacking scenarios.

We start by characterizing $\mathcal{X}^0$ for all possible input-output data generated by the system under a sparse observability condition.

Proposition 2. *Consider system* (1) *with s sensors under attack. Let x_{true} denote the true but unknown initial state. If the system is s-sparse observable, then for any attack signal $\{e(\tau)\}_{0 \leq \tau \leq t}$ assigned by the attacker, the set of all plausible initial states $\mathcal{X}^0$ is a finite set. Moreover, when $p > 2s$, the set of plausible initial states $\mathcal{X}^0$ satisfies*

$$\mathcal{X}^0 \subset \{x_{\text{true}}\} + \bigcup_{\Lambda \in \mathbb{C}_p^{p-2s}} \ker(\mathcal{O}_\Lambda). \tag{9}$$

Proof. Since (A, C) is s-sparse observable, $\mathcal{O}_\Gamma$ is full column rank for any $\Gamma \in \mathbb{C}_p^{p-s}$. Therefore, each set $\mathcal{X}^{0,\Gamma}$ is either a singleton or an empty set. From (4), $\mathcal{X}^0 = \bigcup_{\Gamma \in \mathbb{C}_p^{p-s}} \mathcal{X}^{0,\Gamma}$ is a union of finite sets, so it is finite.

Now consider the case $p > 2s$. Because we can write $x = x_{\text{true}} + (x - x_{\text{true}})$, we will show that $x - x_{\text{true}} \in \bigcup_{\Lambda \in \mathbb{C}_p^{p-2s}} \ker(\mathcal{O}_\Lambda)$ for any $x \in \mathcal{X}^0$. Since $\mathcal{X}^{0,\Gamma}$ is either an empty set or a singleton, there exists $\Gamma_0, \Gamma \in \mathbb{C}_p^{p-s}$ with

$$\widetilde{Y}_i = \begin{bmatrix} y_i(0) \\ y_i(1) \\ \vdots \\ y_i(t) \end{bmatrix}, \ E_i = \begin{bmatrix} e_i(0) \\ e_i(1) \\ \vdots \\ e_i(t) \end{bmatrix}, \ U = \begin{bmatrix} u(0) \\ u(1) \\ \vdots \\ u(t-1) \\ 0 \end{bmatrix}, \ \mathcal{O}_i = \begin{bmatrix} C_i \\ C_i A \\ \vdots \\ C_i A^t \end{bmatrix}, \ F_i = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ C_i B & 0 & \cdots & 0 & 0 \\ \vdots & & \ddots & & 0 \\ C_i A^{t-1} B & C_i A^{t-2} B & \cdots & C_i B & 0 \end{bmatrix} \tag{5}$$

$\mathcal{O}_{\Gamma_0} x_{\text{true}} = Y_{\Gamma_0}$ and $\mathcal{O}_{\Gamma} x = Y_{\Gamma}$. We note the following for Γ_0, Γ from $\mathbb{C}_p^{p-s}$:

$$p - 2s \leq |\Gamma_0 \cap \Gamma| \leq p - s - 1 \tag{10}$$

because the two sets each have $p-s$ elements out of p total elements, and they must also be distinct. The lower bound implies the intersection $\Gamma_0 \cap \Gamma \neq \emptyset$ is nonempty. Therefore, $\mathcal{O}_{\Gamma_0 \cap \Gamma}$ and $Y_{\Gamma_0 \cap \Gamma}$ are well-defined, and:

$$\mathcal{O}_{\Gamma_0 \cap \Gamma} x_{\text{true}} = Y_{\Gamma_0 \cap \Gamma}, \ \mathcal{O}_{\Gamma_0 \cap \Gamma} x = Y_{\Gamma_0 \cap \Gamma}.$$

Subtracting the above two equations yields:

$$x - x_{\text{true}} \in \ker(\mathcal{O}_{\Gamma_0 \cap \Gamma}) \subseteq \bigcup_{\Gamma_1, \Gamma_2 \in \mathbb{C}_p^{p-s}} \ker(\mathcal{O}_{\Gamma_1 \cap \Gamma_2}). \tag{11}$$

From (10), we only need to consider $\Lambda = \Gamma_1 \cap \Gamma_2$ such that $p - 2s \leq |\Lambda| \leq p - (s+1)$:

$$x - x_{\text{true}} \in \bigcup_{k=p-2s}^{p-(s+1)} \left(\bigcup_{\Lambda \in \mathbb{C}_p^k} \ker(\mathcal{O}_\Lambda) \right).$$

One can verify that for any combinations $\bar{\Lambda}, \Lambda \subseteq [p]$, if $\Lambda \subseteq \bar{\Lambda}$, then $\ker(\mathcal{O}_{\bar{\Lambda}}) \subseteq \ker(\mathcal{O}_\Lambda)$. Thus, we know $\bigcup_{\Lambda \in \mathbb{C}_p^{p-(s+1)}} \ker(\mathcal{O}_\Lambda) \subseteq \ldots \subseteq \bigcup_{\Lambda \in \mathbb{C}_p^{p-2s}} \ker(\mathcal{O}_\Lambda)$. This further implies that (9) holds. □

Proposition 2 characterizes the set $\mathcal{X}^0$ when there are sensor redundancies. The result shows that even under severe sensor attacks, s-sparse observability still ensures a finite number of plausible states, and with enough sensors, the attackers can only confuse the system with plausible initial states contained within the set given by (9). We note that the existing result on $2s$-sparse observability follows directly from our characterization.

Corollary 2 ([9]). *When system* (1) *is* $2s$*-sparse observable, we have* $\mathcal{X}^0 = \{x_{\text{true}}\}$.

Proof. Note that $2s$-sparse observable implies $p > 2s$. Since $\ker(\mathcal{O}_\Lambda) = \{0\}, \ \forall \Lambda \in \mathbb{C}_p^{p-2s}$, the result follows from (9). □

Corollary 3. *When system* (1) *is not* s*-sparse observable, there exists some possible attack signal* $\{e(\tau)\}_{0 \leq \tau \leq t}$ *assigned by the attacker such that* $\mathcal{X}^0$ *is an infinite set.*

Proof. Since (A, C) is not s-sparse observable, we know $\mathcal{O}_\Gamma$ is not full column rank for some $\Gamma \in \mathbb{C}_p^{p-s}$. Denote such an index set by Γ_0. Suppose that sensors in the index set Γ_0 are intact, then $\mathcal{X}^{0,\Gamma_0}$ is an affine subspace in $\mathbb{R}^n$ with a dimension of at least 1. From (4), $\mathcal{X}^0$ is thus an infinite set. □

We now restrict our discussion to the case when the system is s-sparse observable. When this condition fails, the sensor attack is undetectable[1]. Now we derive conditions on H and q for system safety in all possible attacking scenarios.

[1]By attack detection, we mean that whether, for any attacks on s sensors, the system can tell apart the case where every sensors are intact and the case where there are sensor attacks in the system. An attack detection on s sensors is possible if and only if the system is s-sparse observable. See [23, Theorem 16.1] for more details.

Theorem 1. *Consider system* (1) *with* s *sensors under attack. Assume that the system is* s*-sparse observable and* $p > 2s$*. If the following conditions hold:*

i. for any $\Lambda \in \mathbb{C}_p^{p-2s}$, $\ker(\mathcal{O}_\Lambda) \subseteq \ker \left(\begin{bmatrix} H \\ HA \\ \vdots \\ HA^{n-1} \end{bmatrix} \right)$;

ii. $h(x) = Hx + q$ *is a control barrier function.*

Then there exist maps $k_t : \mathbb{R}^{m(t-1)+pt} \to \mathbb{R}^m$ *such that the system under the feedback* $u(t) = k_t(\mathcal{D}_t)$ *satisfies the following implications:*

$$\begin{aligned} x(n-1) \in \mathcal{C} &\implies x(t) \in \mathcal{C}, \ \forall t \geq n, \\ x(n-1) \notin \mathcal{C} &\implies \lim_{\tau \to \infty} \|x(\tau)\|_{\mathcal{C}} \to 0. \end{aligned} \tag{12}$$

Proof. We will show that at each time $t \geq n-1$, $u(t)$ can be found to satisfy the CBF condition for all plausible states:

$$H(Ax + Bu(t)) + q \geq (1 - \gamma)(Hx + q), \ \forall x \in \mathcal{X}^t. \tag{13}$$

Using (9) and Corollary 1 to conservatively bound $\mathcal{X}^t$, it is sufficient to show that $u(t)$ satisfies the following condition:

$$\begin{aligned} HBu(t) + H(A - (1-\gamma)I)(x + A^t \cup_{\Lambda \in \mathbb{C}_p^{p-2s}} \ker(\mathcal{O}_\Lambda)) \\ + \gamma q \subseteq \mathbb{R}^l_{\geq 0} \end{aligned} \tag{14}$$

for an arbitrary $x \in \mathcal{X}^t$. Under Condition i, we know for any $\Lambda \in \mathbb{C}_p^{p-s}$, any $v \in \ker(\mathcal{O}_\Lambda)$, $Hv = HAv = \ldots = HA^{n-1}v = 0$. Using the Cayley–Hamilton Theorem [24, Theorem 6.1], we further deduce that $HA^\tau v = 0$ for any $\tau \geq n$. Thus, the condition simplifies to:

$$HBu(t) + H(A - (1-\gamma)I)x + \gamma q \geq 0$$

for an arbitrary $x \in \mathcal{X}^t$ (as opposed to for all $x \in \mathcal{X}^t$ as in (13)). The existence of such an input u is guaranteed by Condition ii. Thus, the feedback map k_t can be given by, for example, solving the following quadratic program

$$\begin{aligned} k_t(\mathcal{D}_t) = \arg\min_u & \|u\|^2 \\ \text{s.t. } & HBu + H(A - (1-\gamma)I)x + \gamma q \geq 0, \end{aligned}$$

where the state x is one plausible state in $\mathcal{X}^t$ given the input-output data $\mathcal{D}_t$. The properties in (12) thus follow from CBF theory [14], [22]. □

Theorem 1 provides conditions on H and q such that the system can be rendered safe. An example using these conditions to design a safe set is given in the simulation section.

We note that while reasoning about system safety offline is useful, the proposed conditions might be too pessimistic in practice since we consider the worst-case attacking scenario. Theoretically, under the s-sparse observability condition, in order to guarantee the system safety starting from $\mathcal{X}^0$, which is a finite set, we have to take into account the set $\{x_{\text{true}}\} + \bigcup_{\Lambda \in \mathbb{C}_p^{p-2s}} \ker(\mathcal{O}_\Lambda)$, which is an infinite set in general. Practically, when considering a robotic system, we argue it is not restrictive to assume that the robot starts in a safe region (though precision location may be unknown), safely collects some input-output data (e.g., by staying still at

its current position), and does not begin its desired task until safety requirements are fulfilled. In the following subsection, we will introduce such requirements, which can be verified for the system after collecting measurements of the first few steps.

C. Online phase safe control design

In this section we derive conditions on the system safety after collecting the input-output data for the first $t \geq n$ steps. Recall that the state uncertainty with the past input-output data is characterized in Proposition 1. We note that s-sparse observability is no longer required in this subsection.

To simplify our analysis, we classify different types of combinations Γ based on the dimensions of $\mathcal{X}^{0,\Gamma}$. Specifically, let

$$\mathbb{C}_p^{p-s} = \mathbb{C}_\emptyset \cup \mathbb{C}_{pt} \cup \mathbb{C}_{sb},$$

where $\mathbb{C}_\emptyset, \mathbb{C}_{pt}, \mathbb{C}_{sb}$ are the sets of combinations that correspond to $\mathcal{X}^{0,\Gamma}$ being an empty set, a singleton, and an affine subspace in $\mathbb{R}^n$, respectively. Let $x^{0,\Gamma}$ satisfy $\mathcal{O}_\Gamma x^{0,\Gamma} = Y_\Gamma$ for $\Gamma \in \mathbb{C}_{pt} \cup \mathbb{C}_{sb}$. Then, from (8), we have

$$\begin{aligned} \mathcal{X}^{t,\Gamma} &= \{x^{t,\Gamma}\} && \text{for } \Gamma \in \mathbb{C}_{pt}, \\ \mathcal{X}^{t,\Gamma} &= \{x^{t,\Gamma}\} + A^t(\ker(\mathcal{O}_\Gamma)) && \text{for } \Gamma \in \mathbb{C}_{sb}, \end{aligned} \tag{15}$$

with $x^{t,\Gamma} = A^t x^{0,\Gamma} + A^{t-1}Bu(0) + A^{t-2}Bu(1) + ... + Bu(t-1)$ propagated from $x^{0,\Gamma}$.

Lemma 1. *The set $\mathcal{X}^t \subseteq \mathcal{C}$ is contained within the safe set if and only if the following conditions hold:*

i. $\ker(\mathcal{O}_\Gamma) \subseteq \ker(HA^t)$ *for all* $\Gamma \in \mathbb{C}_{sb}$*;*

ii. there exists $x^{t,\Gamma} \in \mathcal{X}^{t,\Gamma}$ *satisfying* $Hx^{t,\Gamma} + q \geq 0$ *for each* $\Gamma \in \mathbb{C}_{pt} \cup \mathbb{C}_{sb}$.

Proof. Necessity: Suppose there exists $\Gamma \in \mathbb{C}_{sb}$ such that $\ker(\mathcal{O}_\Gamma) \not\subseteq \ker(HA^t)$, then there exists $v \in \ker(\mathcal{O}_\Gamma)$ such that $HA^t v \neq 0$. However, in view of (15), $x(t) + kA^t v \in \mathcal{X}^t$, for any $k \in \mathbb{R}$, is a plausible state based on data, and a k exists such that $[kHA^t v + Hx(t) + q]_i \not\geq 0$ for some row i, i.e., $\mathcal{X}^t \not\subseteq \mathcal{C}$. This contradiction proves $\ker(\mathcal{O}_\Gamma) \subseteq \ker(HA^t)$ for all $\Gamma \in \mathbb{C}_{sb}$. The second condition follows directly from the definition of the safe set.

Sufficiency: We must show that all $x \in \mathcal{X}^{t,\Gamma}$ for which $x \neq x^{t,\Gamma}$ are also in the safe set, for $\Gamma \in \mathbb{C}_{sb}$ when $\mathcal{X}^{t,\Gamma}$ is not a singleton. For such a state x, one calculates that $h(x) - h(x^{t,\Gamma}) = H(x - x^{t,\Gamma}) = HA^t v$ for some $v \in \ker(\mathcal{O}_\Gamma)$, using (15). Based on the first condition, $h(x) - h(x^{t,\Gamma}) = 0$. Thus x also belongs to $\mathcal{C}$. This completes the proof. □

We are now in place to derive the safety condition on $u(t)$. The naive discrete-time CBF condition on u is

$$H(Ax + Bu(t)) + q \geq (1-\gamma)(Hx + q), \forall x \in \mathcal{X}^t, \tag{16}$$

which may consist of infinitely many linear constraints, which makes it intractable to solve using the standard quadratic-program-based feedback design.

Theorem 2. *Under the premises of Proposition 1, the CBF condition in* (16) *is equivalent to:*

$$\ker(\mathcal{O}_\Gamma) \subseteq \ker(HA^{t+1} - (1-\gamma)HA^t), \ \forall \Gamma \in \mathbb{C}_{sb}, \tag{17}$$

and there exists a plausible state $x^{t,\Gamma} \in \mathcal{X}^{t,\Gamma}$ such that there exists an input $u \in \mathbb{R}^m$ that satisfies:

$$H(Ax^{t,\Gamma} + Bu + q \geq (1-\gamma)(Hx^{t,\Gamma} + q), \tag{18}$$

for each $\Gamma \in \mathbb{C}_{pt} \cup \mathbb{C}_{sb}$. Moreover, when $\mathcal{X}^0, ..., \mathcal{X}^{n-1} \subseteq \mathcal{C}$, the condition in (17) *holds.*

Proof. In view of Corollary 1 and (15), we rewrite the condition in (16) as

$$HBu(t) + H(A - (1-\gamma)I)x + \gamma q \geq 0, \quad \forall x \in \mathcal{X}^{t,\Gamma}, \Gamma \in \mathbb{C}_{pt}. \tag{19a}$$

$$HBu(t) + H(A - (1-\gamma)I)\mathcal{X}^{t,\Gamma} + \gamma q \subseteq \mathbb{R}^l_{\geq 0}, \quad \forall \Gamma \in \mathbb{C}_{sb}. \tag{19b}$$

For $\Gamma \in \mathbb{C}_{pt}$, the set $\mathcal{X}^{t,\Gamma}$ is a singleton, and (19a) is equivalent to (18). For $\Gamma \in \mathbb{C}_{sb}$, one can verify that (19b) holds if and only if conditions (17) and (18) hold by following a proof similar to that of Lemma 1.

From Lemma 1, $\mathcal{X}^0, ..., \mathcal{X}^{n-1} \subseteq \mathcal{C}$ implies $\ker(\mathcal{O}_\Gamma) \subseteq \ker(HA^\tau)$ for $\tau = 0, 1, ..., n-1$ and for any $\Gamma \in \mathbb{C}_{sb}$, i.e., $Hv = HAv = ... = HA^{n-1}v = 0$ for any $v \in \ker(\mathcal{O}_\Gamma)$. From Cayley–Hamilton Theorem [24, Theorem 6.1], we further know $HA^\tau v = 0$ for $\tau \geq n$. Thus, the condition in (17) is satisfied. □

Remark 1. The condition on $u(t)$ in (18) is finite and linear, and thus can be implemented as constraints in quadratic program-based control designs.

We now have the following results regarding system safety.

Theorem 3. *Under the premises of Proposition 1, if $\mathcal{X}^0, \ldots, \mathcal{X}^{n-1} \subseteq \mathcal{C}$, and if the condition in* (18) *holds at each time step $\tau \geq t$ with a control sequence $\{u(\tau)\}_{\tau \geq t}$, then the two implications in* (12) *hold for the system.*

Proof. From Theorem 2, we know that the CBF condition (16) is fulfilled for all states in $\mathcal{X}^\tau$ for all $\tau \geq t$. As the CBF condition is fulfilled along the system trajectory, we thus know that the implications in (12) hold following CBF theory [14], [22]. □

Theorems 2 and 3 require the conditions (17) and (18), which may be difficult to verify. Here, we provide alternative, albeit only sufficient, conditions that are easier to verify.

Corollary 4. *Sufficient conditions for the existence of an input $u \in \mathbb{R}^m$ such that conditions* (17) *and* (18) *hold are:*

$$\forall \Gamma \in \mathbb{C}_{sb}, \ \exists M_\Gamma \in \mathbb{R}^{l \times (p-s)} \text{ s.t. } H = M_\Gamma C_\Gamma, \tag{20}$$

$$\text{and } HB \in \mathbb{R}^{l \times n} \text{ is full row rank.} \tag{21}$$

Proof. We first show that (20) implies condition (17). For any $v \in \ker(\mathcal{O}_\Gamma)$, we have $\mathcal{O}_\Gamma v = 0$ implies $C_\Gamma v = C_\Gamma Av = \cdots = C_\Gamma A^{n-1}v = 0$ by definition. Using Cayley-Hamilton Theorem [24, Theorem 6.1], we further deduce $C_\Gamma A^{t+1}v = C_\Gamma A^t v = 0$. Therefore, $(HA^{t+1} - (1-\gamma)HA^t)v = M_\Gamma C_\Gamma (A^{t+1} - (1-\gamma)A^t)v = 0$. That is,

$v \in \ker(HA^{t+1} - (1-\gamma)HA^t)$. Thus the equality in (17) holds for any $\Gamma \in \mathbb{C}_{sb}$.

We now show that (21) implies the existence of an input u such that (18) holds. Note that there are finite number of sets $\Gamma \in \mathbb{C}_{pt} \cup \mathbb{C}_{sb}$, i.e. $|\mathbb{C}_{pt} \cup \mathbb{C}_{sb}| = k \in \mathbb{N}$, so we seek to show there exists $u \in \mathbb{R}^m$:

$$HBu + z \geq 0 \tag{22}$$

where $z = (z_1, ..., z_k)$ and

$$z_i := \min_{\Gamma \in \mathbb{C}_{pt} \cup \mathbb{C}_{sb}} [H(A - (1-\gamma)I)A^t x^{t,\Gamma} + \gamma q]_i$$

When (21) holds, one feasible solution to (22) is $u = B^\top H^\top (HBB^\top H^\top)^{-1} z$, which also fulfills the condition in (18). □

IV. A 2D-VEHICLE EXAMPLE

Consider an omnidirectional vehicle on a 2D-plane with continuous-time dynamics

$$\dot{x} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -0.2 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -0.2 \end{bmatrix} x + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} u \tag{23}$$

where $x = (x_1, x_2, x_3, x_4), u = (u_1, u_2)$, x_1, x_3 are the x, y-axes positions, x_2, x_4 the respective velocities, and u_1, u_2 the respective acceleration-level inputs. For digital implementation of our controller, this continuous-time system is discretized in time using zero-order hold method with sampling time 0.01s. The system output is given by

$$y = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} x + e, \tag{24}$$

where recall e represents the attacking signal.

Offline safety guarantee: One verifies that this system is 1-sparse observable. When only 1 sensor attack exists, following Proposition 2, we know the set of plausible initial states

$$\mathcal{X}^0 \subseteq \{x_{\text{true}}\} + \text{span}\left(\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}\right) \cup \text{span}\left(\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}\right).$$

From Theorem 1, one possible safe region is given by $\mathcal{C} = \{x : h(x) = Hx + q \geq 0\}$, where

$$H = M \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

for any M with proper dimensions and q such that the function h is a CBF. Theorem 1 guarantees that if the system state is inside the safe set at the 4th step, then the system can be rendered safe in all possible attacking scenarios. One example choice is $\mathcal{C} = \{x : -4 \leq x_i \leq 4, i = 2, 4\}$.

Online safe control: In this case, the safe set is chosen as

$$\mathcal{C} = \{x : -4 \leq x_i \leq 4, i = 1, 2, 3, 4\}. \tag{25}$$

Here, we consider that sensors $1, 3, 5$ are under attack and the attacker intends to confuse the system with a fake initial state $x_{\text{fake}} = (2, 2, 2, 1)$ while the actual initial state $x_{\text{true}} = (1, 1, 1, 1)$. Measurements of attacked sensors are computed based on the fake trajectory starting from x_{fake} according to the system dynamics. In this simulation, Gaussian noises from $\mathcal{N}(0, 0.01^2)$ are added to all the measurements.

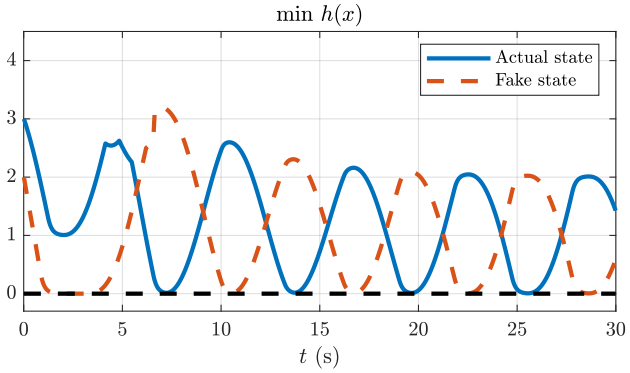

Fig. 1: CBF evolution over time

For online state estimation, we take a brute-force approach of looking at all possible sensor combinations in $\mathbb{C}_p^{p-s}$ and determining whether the sensors are intact. This is done by computing the least squares solution to the linear equality $\mathcal{O}_\Gamma x = Y_\Gamma$, and empirically checking the bound of the matching error. In this simulation, we set the error threshold to 0.001. We note that even though in the theoretical development we consider all the measurements from time steps 0 to t for determine the state at time step t, it is not necessary nor practical to store all those measurements for secure state reconstruction. Instead, we will use measurements at $t-3, t-2, t-1, t$ to determine the plausible states at time step t. The plausible states at time step $t-3$ obtained from (4) are then propagated to the plausible states at time step t according to (8).

According to Theorem 2, we apply the following online safe controller

$$\begin{aligned} u(t) = \arg\min & \|u - u_{\text{nom}}\| \\ & H(Ax^{t,\Gamma} + Bu(t)) + q \geq (1-\gamma)(Hx^{t,\Gamma} + q), \\ & \forall \Gamma \in \mathbb{C}_{pt} \cup \mathbb{C}_{sb} \end{aligned} \tag{26}$$

together with a one-time checking mechanism on whether the plausible states at the first 4 steps lie within the safe set $\mathcal{C}$. From Theorem 3, system safety is guaranteed if this QP is always feasible. A nominal input at time step τ is chosen as $u_{\text{nom}}(\tau) = (\sin(0.01\tau), \cos(0.01\tau))$. This control signal is implemented for the first three steps and then filtered out safely by the QP in (26), where γ is chosen to be 0.05.

During implementation, the online state estimator finds 4 plausible initial states $(1,1,1,1)$, $(1,1,2,1)$, $(2,2,1,1)$, $(2,2,2,1)$. We note that the system cannot distinguish the actual state from the other three fake states.

We observe from Fig. 1 that the actual trajectory respects the safety constraint for all time. The fake trajectory also fulfills the safety constraint. This, however, is not always the case. For example, when the attacker chooses $x_{\text{fake}} = (2,2,2,2)$, the corresponding fake trajectory does not fulfill the safety requirement. This is due to the fact that by attacking sensors $1, 3$, and 5, the attacker can not confuse

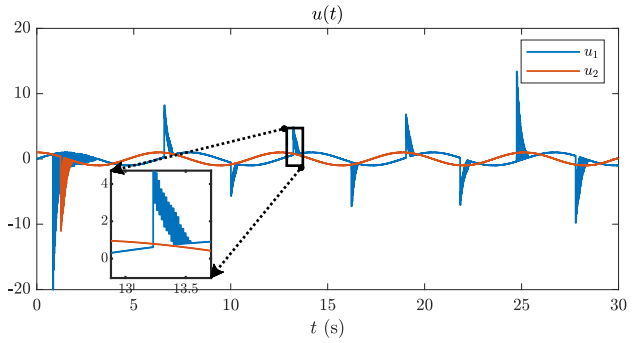

Fig. 2: Safe control input over time

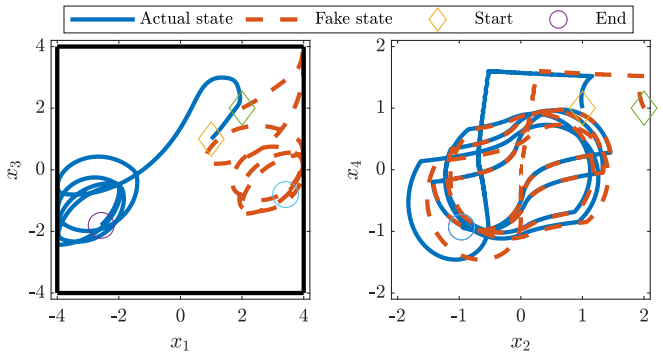

Fig. 3: Actual trajectory v.s. fake trajectory

the system about the velocity along y-axis x_4, and $x_4 = 1$ for all the plausible initial states $x \in \mathcal{X}^0$.

Figure 2 demonstrates that the safe control input closely resembles the nominal one, and only modifies the nominal control when necessary. One can also see from Fig. 3 that although the velocity responses of the actual and the fake trajectories share some similarities, the positional movement is very different. Our proposed safe controller correctly constrains all possible trajectories inside the safe region (the square enclosed by black lines).

V. Conclusions

In this work, we have provided conditions that guarantee safety for discrete-time linear systems under severe sensor attacks. We consider a scenario where the secure state reconstruction problem may have non-unique solutions. We first provide an exact characterization of these plausible states. We then derive conditions for designing an offline safe set, which guarantees system safety for all possible sensor attacks under a mild sparse observability condition. When the system is deployed online with measurement data for the first few steps available, we have proposed a quadratic program-based safe control scheme. We show that system safety hinges on a finite number of control barrier function conditions, and on a kernel condition that is related to the system dynamics and the number of attacked sensors. A numerical example of a 2-D vehicle illustrates theoretical results.

References

[1] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Security & Privacy*, vol. 9, no. 3, pp. 49–51, 2011.

[2] Y. Shoukry, P. Martin, P. Tabuada, and M. Srivastava, "Non-invasive spoofing attacks for anti-lock braking systems," in *Cryptographic Hardware and Embedded Systems-CHES 2013: 15th International Workshop, Santa Barbara, CA, USA, August 20-23, 2013. Proceedings 15*. Springer, 2013, pp. 55–72.

[3] Y. Tu, Z. Lin, I. Lee, and X. Hei, "Injected and delivered: Fabricating implicit control over actuation systems by spoofing inertial sensors," in *27th USENIX security symposium (USENIX Security 18)*, 2018, pp. 1545–1562.

[4] C. De Persis and P. Tesi, "Input-to-state stabilizing control under denial-of-service," *IEEE Transactions on Automatic Control*, vol. 60, no. 11, pp. 2930–2944, 2015.

[5] M. Zhu and S. Martinez, "On the performance analysis of resilient networked control systems under replay attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 804–808, 2013.

[6] R. S. Smith, "Covert misappropriation of networked control systems: Presenting a feedback structure," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 82–92, 2015.

[7] Y. Mo, E. Garone, A. Casavola, and B. Sinopoli, "False data injection attacks against state estimation in wireless sensor networks," in *49th IEEE Conference on Decision and Control (CDC)*. IEEE, 2010, pp. 5967–5972.

[8] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic control*, vol. 59, no. 6, pp. 1454–1467, 2014.

[9] Y. Shoukry and P. Tabuada, "Event-triggered state observers for sparse sensor noise/attacks," *IEEE Transactions on Automatic Control*, vol. 61, no. 8, pp. 2079–2091, 2015.

[10] M. S. Chong, M. Wakaiki, and J. P. Hespanha, "Observability of linear systems under adversarial attacks," in *2015 American Control Conference (ACC)*. IEEE, 2015, pp. 2439–2444.

[11] Y. Mao, A. Mitra, S. Sundaram, and P. Tabuada, "On the computational complexity of the secure state-reconstruction problem," *Automatica*, vol. 136, p. 110083, 2022.

[12] Y. Lin, M. S. Chong, and C. Murguia, "Plug-and-play secondary control for safety of LTI systems under attacks," *arXiv preprint arXiv:2212.00593*, 2022.

[13] H. Zhang, Z. Li, and A. Clark, "Safe control for nonlinear systems under faults and attacks via control barrier functions," *arXiv preprint arXiv:2207.05146*, 2022.

[14] A. D. Ames, X. Xu, J. W. Grizzle, and P. Tabuada, "Control barrier function based quadratic programs for safety critical systems," *IEEE Transactions on Automatic Control*, vol. 62, no. 8, pp. 3861–3876, 2016.

[15] S. Kolathaya and A. D. Ames, "Input-to-state safety with control barrier functions," *IEEE control systems letters*, vol. 3, no. 1, pp. 108–113, 2018.

[16] D. R. Agrawal and D. Panagou, "Safe and robust observer-controller synthesis using control barrier functions," *IEEE Control Systems Letters*, vol. 7, pp. 127–132, 2022.

[17] S. Dean, A. Taylor, R. Cosner, B. Recht, and A. Ames, "Guaranteeing safety of learned perception modules via measurement-robust control barrier functions," in *Conference on Robot Learning*. PMLR, 2021, pp. 654–670.

[18] A. Singletary, M. Ahmadi, and A. D. Ames, "Safe control for nonlinear systems with stochastic uncertainty via risk control barrier functions," *IEEE Control Systems Letters*, vol. 7, pp. 349–354, 2022.

[19] M. Vahs, C. Pek, and J. Tumova, "Belief control barrier functions for risk-aware control," *IEEE Robotics and Automation Letters*, 2023.

[20] R. K. Cosner, P. Culbertson, A. J. Taylor, and A. D. Ames, "Robust safety under stochastic uncertainty with discrete-time control barrier functions," *arXiv preprint arXiv:2302.07469*, 2023.

[21] B. Zhong, S. Liu, M. Caccamo, and M. Zamani, "Secure-by-construction controller synthesis via control barrier functions," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 239–245, 2023.

[22] A. Agrawal and K. Sreenath, "Discrete control barrier functions for safety-critical control of discrete systems with application to bipedal robot navigation." in *Robotics: Science and Systems*, vol. 13. Cambridge, MA, USA, 2017, pp. 1–10.

[23] S. Diggavi and P. Tabuada, "A coding theoretic view of secure state reconstruction," *Modeling and Design of Secure Internet of Things*, pp. 357–369, 2020.

[24] J. P. Hespanha, *Linear systems theory*. Princeton university press, 2018.