Enhancing Sensor Attack Detection and Mitigating Sensor Compromise Impact in a Switching-Based Moving Target Defense

Anas Alhashimi[1,2], Thomas Nolte[1], Alessandro V. Papadopoulos[1]

***Abstract*— This study is based on a Moving Target Defence (MTD) algorithm designed to introduce uncertainty into the controller and another layer of uncertainty to intrusion detection. This randomness complicates the adversary's attempts to craft stealthy attacks while concurrently minimizing the impact of false-data injection attacks. Leveraging concepts from state observer design, the method establishes an optimization framework to determine the parameters of the random signals. These signals are strategically tuned to increase the detectability of stealthy attacks while reducing the deviation resulting from false data injection attempts. We propose here to use two different state observers and two associated MTD algorithms. The first one optimizes the parameters of the random signals to reduce the deviation resulting from false data injection attempts and maintain the stability of the closed-loop system with the desired level of performance. In contrast, the second one optimizes the parameters of the random signals to increase the detectability of stealthy attacks. Dividing the optimization problem into two separate optimization processes simplifies the search process and makes it possible to have higher values of the detection cost function. To illustrate the effectiveness of our approach, we present a case study involving a generic linear time-invariant system and compare the results with a recently published algorithm.**

I. Introduction

The term Cyber-Physical System (CPS) refers to an integrated platform that encompasses an outer physical layer containing sensors and actuators alongside a communication layer and a control layer (see Figure 1). Such systems have many applications, including safety monitoring, smart grids, healthcare, infrastructure, and transportation.

CPS attacks in the real world can have far-reaching consequences, including damage to critical infrastructure, public safety risks, and economic effects. Examples of such attacks are: Maroochy Water Services Hack [1] releasing sewage into public areas in Australia, Stuxnet Worm [2] caused physical damage and disrupted Iran's nuclear program, NotPetya [3] caused a cumulative financial loss of 10 billion dollars to some of the major industries worldwide, and TRITON attack [4] caused shutdown to petrochemical processing plant in Saudi Arabia, etc. Therefore, there is an ongoing need for robust cyber-security measures to protect critical infrastructure and control systems from such threats.

Cyber systems' inherent static nature grants attackers the advantage of time. Thankfully, a novel strategy known as Moving Target Defence (MTD) has arisen as a promising solution to address this issue [5]. In response to this vulnerability, MTD has arisen as a strategic approach aimed at introducing unpredictability to the state and operation of a system. This serves the dual purpose of thwarting adversaries from achieving predictability in the outcomes of their attacks and enhancing the possibility of detecting stealthy intrusion attempts.

This research is motivated by the switching-based MTD algorithm against sensor attack presented in the recent work [6], [7], and modifies the *attack detection* part. It assumes the adversary knows system dynamics and attack-detection strategy and has access to all control inputs and all sensor readings. The original algorithm has the properties of enhancing the capability to identify highly stealthy attacks and minimizing the impact of a sensor compromise in a controlled manner by solving an optimization problem to design the parameters to minimize the impact of attacks. We aim to improve detection performance by increasing the margin between the detection statistics and the threshold values, reducing the effects of sensor attacks on system states, and maintaining the system's transient response to ensure earlier attack detection. Our proposed method is based on two independent optimizations, one for enhancing the detection of stealthy attacks and the second for minimizing the impact of a sensor compromise. To facilitate this separation, another observer is added to the attack detection system with its own MTD algorithm. An interesting key feature of this method is that it can be designed to preserve the stability of the original control system. The performance of the resulting system with the first MTD algorithm is specified by design, while the second MTD does not affect the impact of attacks, thanks to the separation principle.

A. Related work

Multiple strategies are employed to enhance the security and resilience of systems using MTD [8], [9]. In the context of smart grids, a key approach involves modifying the system's physical topology to improve state estimation and thereby uncover potential false-data-injection attacks [10], [11]. Another strategy revolves around the introduction of external states that are linked to the state of the control system. Sensors are utilized to monitor and measure these external states. The sensor-based measurements of external states add a layer of security and situational awareness, making it more difficult for adversaries to exploit vulnerabilities [12], [13]. Embedding a watermark within the control signal, and if the anomaly detection system cannot detect it within the sensor readings, it initiates an alert [14]. Combining watermarking

This work was supported by the Swedish Research Council (VR) via the project "Pervasive Self-Optimizing Computing Infrastructure (PSI)", and by the Knowledge Foundation (KKS) via the project FIESTA.

[1] Mälardalen University, Västerås, Sweden. [2]University of Baghdad, Baghdad, Iraq.

and MTD [15]. Increase the uncertainty of the system by randomly switching among several controllers [16]. Injecting random noise into the controller makes it considerably more challenging for potential adversaries to estimate the output accurately [17]. Using IoT is a powerful method to replicate sensor data. Following replication, advanced estimation theory tools are applied to isolate sensors that may have been compromised [18]. Another approach involves duplicating sensory and control signals, transmitting them randomly through separate communication channels, and then randomly selecting one of the duplicated signals to deliver to its intended destination [19].

A time-delayed switching model and observer-based security control scheme presented by [20] for CPSs. Using switching in securing control systems has been proposed earlier in [21], the authors proposed a game theoretic approach to estimate a binary random variable based on sensor measurements that a cyber-attacker may have corrupted. A switching-based MTD is presented in [22] and [23] to detect false-data-injection with an Event-Triggered mechanism to reduce the cost of data transmission. By leveraging techniques from observer design and switched systems, Giraldo et al. [7] developed an MTD algorithm and defined an optimization problem aimed at determining the probability of switching signals that enhance the detectability of stealth attacks, all while reducing the impact of false-data-injection attacks.

We adopted the switching-based MTD in [7]; it can be implemented in two distinct ways: In the first method, a MTD random signal is introduced into the sensor outputs of the physical system, and an identical one into the observer. This setup increases uncertainty for both the system and the controller, although it requires synchronization between the physical system and the controller. In the second method, the one under consideration here, the MTD random signal, is exclusively applied to the controller and Intrusion Detection System (IDS). Consequently, increased uncertainty is introduced only to the controller and IDS, leaving the physical process unaffected. Importantly, this approach eliminates the need for synchronization, making it easier to implement and more widely applicable.

B. Contributions

We summarize our contributions here in the following,

- A novel architecture for switching-based MTD system that incorporates a dedicated observer and MTD within the IDS to add another layer of system uncertainty to the adversary.
- Formulate a pair of distinct optimization problems, the first to improve the detection of stealth attacks and the second to reduce attack impact on the system.

C. Manuscript organization

Formulating the problem in Section II, outlining the methods in Section III, designing the MTD in Section IV, presenting the case study in Section V, and summarizing the conclusions in Section VI.

II. Problem Formulation

We examine the control system illustrated in Figure 1, comprising a physical process equipped with sensors and actuators. This system also includes a MTD mechanism, which introduces randomness into the sensor values used by the controller at any given moment. Additionally, an observer-based controller utilizes the available sensor measurements, which have been modified by the MTD mechanism, to estimate system states and generate control commands. Furthermore, an observer-based IDS is integrated into the system. The IDS observer has been modified by another MTD mechanism. The first MTD aims to introduce uncertainty into the controller to limit the adversary's control over the plant. In contrast, the second MTD seeks to introduce an additional layer of uncertainty into the system, thereby increasing potential attackers' difficulty in hiding their actions.

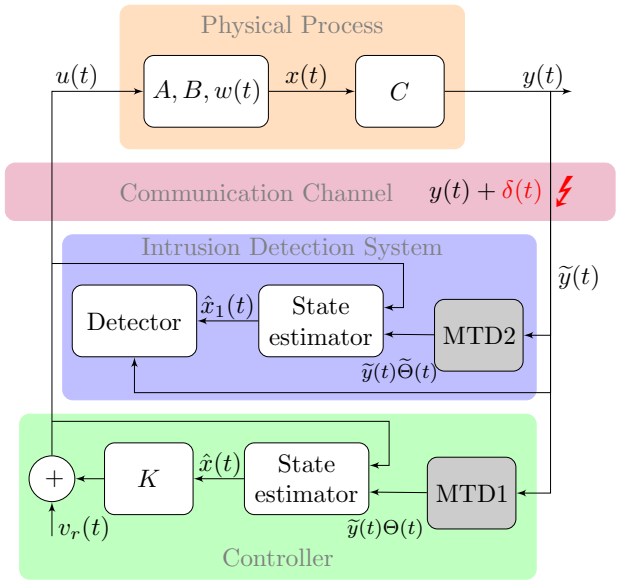

Fig. 1. The system with the proposed MTD mechanism.

Typically, the concept of MTD entails a strategic method intended to inject variability into a system's states and operations. This implies that adopting the MTD strategy aims to thwart attackers from acquiring substantial information. Consequently, MTD measures are implemented before data transmission. However, in Figure 1, the MTD components are incorporated post-output transmission, granting attackers unrestricted access to system outputs. Nevertheless, this setup elevates unpredictability for the controller and IDS, posing challenges to potential attackers.

A. System model

We examine systems that are continuous-time and linear time-invariant, taking the following form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + w(t) \\ y(t) &= Cx(t) + \nu(t) \\ \widetilde{y}(t) &= y(t) + \delta(t) \end{aligned} \tag{1}$$

where

- $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^l$, and $y(t) \in \mathbb{R}^m$ the system states, control input, and sensor outputs of size n, l, and m respectively,
- $\delta(t) \in \mathbb{R}^m$ is the attack vector injected to the sensors,
- $w(t) \sim \mathcal{N}(0, Q)$ and $\nu(t) \sim \mathcal{N}(0, R)$ independent and identically distributed (iid) and $Q, R \succ 0$.

B. State observer with MTD

We propose a state observer given by (2) where the MTD is applied to the innovation term $[\widetilde{y}(t) - C\hat{x}(t)]$

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L\Theta(t)\left[\widetilde{y}(t) - C\hat{x}(t)\right] \quad (2)$$

where $\Theta(t) := \text{diag}(\theta_1(t), \theta_2(t), \ldots, \theta_m(t))$ is a diagonal matrix of independent binary random variables $\theta_i(t) \sim \mathcal{B}(p_i)$ (random variable drawn from a Bernoulli distribution such that $\theta_i(t) = 1$ with probability p_i and zero otherwise). The variables θ_i are assumed to be piecewise linear and stay constant during the period $(t_i, t_{i+1}]$. The holding time $T_i := t_{i+1} - t_i$ is assumed to be a uniformly distributed random variable, $T_i \sim \mathcal{U}[T_{\min}, T_{\max}]$ such that $T_{\min} \leq T_i \leq T_{\max}$.

Define the state estimation error between the states and the observer estimate, $e(t) := x(t) - \hat{x}(t)$, then

$$\begin{aligned} \dot{e}(t) &= \dot{x}(t) - \dot{\hat{x}}(t) \\ &= (A - L\Theta(t)C)\, e(t) - L\Theta(t)\delta(t) - \\ &\quad L\Theta(t)\nu(t) + w(t). \end{aligned} \quad (3)$$

The observer design transforms into a stabilization challenge, necessitating carefully selecting parameters L and $\Theta(t)$. These selections must ensure that the system in (3) achieves global asymptotic stability when $\delta(t) = 0$ and $\nu(t) = 0$ for all time instances. Lastly, it is assumed that the controllability of the pair (A, B) holds, and we consider an output-feedback controller in the following form:

$$u(t) = -K\hat{x}(t). \quad (4)$$

To simplify the analysis, we assume the system is operated at a steady-state point. Therefore, the states and the input signals $v_r(t)$ are assumed to be zero during the system's normal operation. Therefore, we will neglect $v_r(t)$ during the remaining part of this paper.

C. IDS observer with MTD

We introduce an additional observer exclusively utilized by the IDS to enhance attack detection. In this section, we are studying two generic observers.

first IDS observer

Consider having an observer described by

$$\dot{\hat{x}}_1(t) = A\hat{x}_1(t) + Bu(t) + L\widetilde{\Theta}(t)\left[\widetilde{y}(t) - C\hat{x}_1(t)\right] \quad (5)$$

where $\widetilde{\Theta}(t)$ is defined in a similar way to $\Theta(t)$ but with different parameters, i.e., $\widetilde{\Theta}(t) := \text{diag}(\widetilde{\theta}_1(t), \widetilde{\theta}_2(t), \ldots, \widetilde{\theta}_m(t))$ the diagonal matrix of the second MTD and $\widetilde{\theta}_i(t) \sim \mathcal{B}(\widetilde{P}_i)$.

We define the associated state estimation error as $\widetilde{e}(t) := x(t) - \hat{x}_1(t)$, then we have the error dynamics

$$\dot{\widetilde{e}}(t) = \left[A - L\widetilde{\Theta}(t)C\right]\widetilde{e}(t) - L\widetilde{\Theta}(t)[\delta(t) + \nu(t)] + w(t). \quad (6)$$

Similar to (3), we need to carefully select the parameters L and $\widetilde{\Theta}(t)$ to ensure that the system in (6) achieves global asymptotic stability when $\delta(t) = 0$ and $\nu(t) = 0$ for all time instances.

second IDS observer

Since we already have $\hat{x}(t)$, we may include it in the observer residual. Consider having an observer described by

$$\begin{aligned} \dot{\hat{x}}_1(t) &= A\hat{x}_1(t) + Bu(t) + \\ &\quad L\widetilde{\Theta}(t)\left[\alpha\widetilde{y}(t) - \alpha_2 C\hat{x}(t) - \alpha_1 C\hat{x}_1(t)\right] \end{aligned} \quad (7)$$

where $\alpha = \alpha_1 + \alpha_2$ and $\widetilde{\Theta}(t)$ as in above. The state estimation error will be

$$\begin{aligned} \dot{\widetilde{e}}(t) &= \left[A - \alpha_1 L\widetilde{\Theta}(t)C\right]\widetilde{e}(t) - \alpha_2 L\widetilde{\Theta}(t)Ce(t) \\ &\quad -\alpha L\widetilde{\Theta}(t)[\delta(t) + \nu(t)] + w(t). \end{aligned} \quad (8)$$

Adding the term with $\hat{x}(t)$ makes the second observer's error depend on the first observer's error. Which may not be advisable unless there is a good reason behind it. However, the optimization complexity is not affected if the parameters α_1 and α_2 are set before the optimization, but we need to do the design of the first MTD before the second one as explained in Section IV. The error dynamics in (6) could be seen as a special case of (8) in which $\alpha_1 = 1$ and $\alpha_2 = 0$.

D. Intrusion detection system

We can create an anomaly detection module by utilizing the state estimator described in (2). This module can be designed to compare the estimated sensor readings with the actual sensor readings to identify an attack's occurrence. Consequently, we define the residual as follows:

$$\begin{aligned} r(t) &:= \widetilde{y}(t) - C\hat{x}(t) \\ &= Ce(t) + \delta(t) + \nu(t). \end{aligned} \quad (9)$$

Consider the bad-data detection with the detection statistics

$$h(t) = |r(t)|.$$

An alarm is triggered when any $h_i(t) > \tau_i$ for some fixed detection threshold $\tau_i > 0$ is computed based on a given false-alarm probability and noise statistics. The detection threshold for a normally distributed signal with variance σ^2 and mean μ for a given Probability of False Alarm (P_{FA}) can be expressed using the error function complement (erfc) approximation as follows

$$\tau = \mu + \sigma \cdot \text{erfc}^{-1}(2P_{\text{FA}}). \quad (10)$$

This is the exact detector used in [7], however, we propose here to replace the residual (9) with the alternative residual (11)

Improved detector

Consider the following residual

$$\begin{aligned} \widetilde{r}(t) &= \widetilde{y}(t) - C\hat{x}_1(t) \\ &= C\widetilde{e}(t) + \delta(t) + \nu(t) \end{aligned} \quad (11)$$

and the detection statistics

$$\widetilde{h}(t) = |\widetilde{r}(t)|.$$

Notice that $\hat{x}_1(t)$ could be either from (5) or (7) depending on which observer is being used.

This will significantly improve the detection performance compared to the original algorithm in [7] since the MTD applied on $\hat{x}_1(t)$ observer is optimized to have the best detection performance (both detection sensitivity and speed) while the MTD applied on $\hat{x}(t)$ observer is optimized to stabilizing the system, maintaining the convergence speed, and to reduce the attack effect. Compared to the algorithm in [7] where detection performance, stability, and convergence speed were optimized together in a single optimization, which reduces the cost of the detection performance in favor of the other cost function of state drift, doing separate optimization will result in higher scores for the cost function given that the optimization algorithm is working properly.

III. Methods

Here, we discuss system stability first, then the expressions for attack impact on states and detection statistics for the case of simple fixed attack and then the case of stealth attack. Since the system is assumed to be in a steady state, we used the second norm of the system states to measure attack influence on the system.

A. Stability of the switching dynamical system

let $z(t) := \left[x(t)^T, e(t)^T, \widetilde{e}(t)^T\right]^T$ be an extended state vector and consider the second IDS observer (7), the extended system will be

$$\begin{aligned} \dot{z}(t) &= \begin{bmatrix} A+BK & -BK & 0 \\ 0 & A-L\Theta(t)C & 0 \\ 0 & -\alpha_2 L\widetilde{\Theta}(t)C & A-L\alpha_1\widetilde{\Theta}(t)C \end{bmatrix} \\ & \quad z(t) + \begin{bmatrix} 0 \\ -L\Theta(t) \\ -\alpha L\widetilde{\Theta}(t) \end{bmatrix}\delta(t) + \begin{bmatrix} w(t) \\ 0 \\ 0 \end{bmatrix} \\ y(t) &= \begin{bmatrix} C & 0 & 0 \end{bmatrix} z(t). \end{aligned} \tag{12}$$

Notice that the extended system considering the first IDS observer in (5), can be seen as a special case of (12) and can be obtained by setting $\alpha_1 = 1$ and $\alpha_2 = 0$.

Since K is designed such that $A+BK$ is stable, then the stability of (12) is determined by the stability of the submatrix

$$F(t) := \begin{bmatrix} A-L\Theta(t)C & 0 \\ -\alpha_2 L\widetilde{\Theta}(t)C & A-\alpha_1 L\widetilde{\Theta}(t)C \end{bmatrix} \tag{13}$$

We follow the same stability analysis for switched systems presented in [7] since the extended system (12) has a similar structure. The stability of the switching system is based on the result of the study carried out by Chatterjee et al. [24] and Theorem 3.1 in [7].

We summarize the main point as follows, *The feedback system with stable and unstable sub-systems is globally asymptotic stable almost surely if the probability of the unstable subsystems being active is small*.

B. Fixed sensor attack

Consider the attack given by

$$\delta(t) := \tau \tag{14}$$

for some constant vector τ. The expected values of the residuals will be

$$\bar{r}(t) = C\bar{e}(t) + \tau \tag{15}$$

and

$$\bar{\widetilde{r}}(t) = C\bar{\widetilde{e}}(t) + \tau \tag{16}$$

respectively.

Without MTD, to find the detection statistics, we need to find the equilibrium value for $\widetilde{e}(t)$ and $e(t)$. From (3) we have

$$\lim_{t\to\infty} E\left[e(t)\right] = \lim_{t\to\infty} \bar{e}(t) = \left(A-LC\right)^{-1} L\tau \tag{17}$$

sub in (15), we get the statistics

$$\bar{h}(t) = \left| C\left(A-LC\right)^{-1} L\tau + \tau \right| \tag{18}$$

This is the same bad-data detection statistics in [7]; we'll use it here for comparison purposes only. Now, from (8) we have

$$\lim_{t\to\infty} E\left[\widetilde{e}(t)\right] = \left(A-\alpha_1 LC\right)^{-1} L\left(\alpha\tau + \alpha_2 C\bar{e}(t)\right) \tag{19}$$

sub in (16), we obtain

$$\lim_{t\to\infty} \bar{\widetilde{h}}(t) = \left| C\left(A-\alpha_1 LC\right)^{-1} L\left(\alpha\tau + \alpha_2 C\bar{e}(t)\right) + \tau \right|. \tag{20}$$

To find the state drift, we consider the state's dynamics from (12)

$$\dot{x}(t) = (A+BK)x(t) - BKe(t) + w(t) \tag{21}$$

which gives the following equilibrium expression for the states after substituting $\bar{e}(t)$ from (17)

$$\lim_{t\to\infty} E\left[x(t)\right] = \left(A+BK\right)^{-1} BK\left(A-LC\right)L\tau. \tag{22}$$

Notice that We have exactly the same expression for the attack impact on states independently of which detector we are using (9) or (11) since they are not affecting the state estimation

Now we consider the case With MTD. The expected value for the error $e(t)$ will be

$$\lim_{t\to\infty} \bar{e}(t) = \left(A-LPC\right)^{-1} LP\tau \tag{23}$$

where $E\left[\Theta(t)\right] = P$ and $P := \text{diag}(p_1, p_2, \ldots, p_m)$. This gives the detection statistics

$$\bar{h}(t) = \left| C\left(A-LPC\right)^{-1} LP\tau + \tau \right| \tag{24}$$

and for $\widetilde{e}(t)$

$$\lim_{t\to\infty} E\left[\widetilde{e}(t)\right] = \left(A-\alpha_1 L\widetilde{P}C\right)^{-1} L\widetilde{P}\left(\alpha\tau + \alpha_2 C\bar{e}(t)\right) \tag{25}$$

where $E\left[\widetilde{\Theta}(t)\right] = \widetilde{P}$ and $\widetilde{P} := \text{diag}(\widetilde{p}_1, \widetilde{p}_2, \ldots, \widetilde{p}_m)$.

Which gives the detection statistics

$$\bar{\widetilde{h}}(t) = \left| C\left(A - \alpha_1 L\widetilde{P}C\right)^{-1}\left(\alpha L\widetilde{P}\tau + \alpha_2 L\widetilde{P}C\bar{e}(t)\right) + \tau \right|. \tag{26}$$

The state drift, in this case, can be obtained in a similar way to (22) to be

$$\lim_{t\to\infty} E\left[x(t)\right] = \left(A+BK\right)^{-1} BK\left(A-LPC\right)^{-1} LP\tau \tag{27}$$

C. Special stealth sensor attack

Assume the attacker has an estimator described by

$$\dot{\hat{x}}_a(t) = A\hat{x}_a(t) + Bu(t) + L\left[\widetilde{y}(t) - C\hat{x}_a(t)\right] \tag{28}$$

and consider the attack given by

$$\delta(t) := -C(x(t) - \hat{x}_a(t)) + \tau - \nu(t) \tag{29}$$

for some threshold vector τ. Such an attack could be very powerful when the original matrix A has positive eigenvalues and $\hat{x}_a(t)$ perfectly matches $\hat{x}(t)$. Substitute $\delta(t)$ in (12) results in having the following state transition matrix

$$\begin{bmatrix} A+BK & -BK & 0 \\ 0 & A & 0 \\ 0 & \alpha_1 L\widetilde{\Theta}(t)C & A - \alpha_1 L\widetilde{\Theta}(t)C \end{bmatrix} \tag{30}$$

which will make the whole system unstable.

Given the attacker's estimator $\hat{x}_a(t)$ in (28), we define the error between the system estimation used by the controller and the attacker estimation $s(t) := \hat{x}_1(t) - \hat{x}_a(t)$, then the residual will be

$$\begin{aligned} \widetilde{r}(t) &= \widetilde{y}(t) - C\hat{x}_1(t) \\ &= Cx(t) - C\hat{x}_1(t) - C(x(t) - \hat{x}_a(t)) + \tau \\ &= -Cs(t) + \tau. \end{aligned} \tag{31}$$

In order to find the dynamics of $s(t)$ we rewrite (7) as

$$\begin{aligned} \dot{\hat{x}}_1(t) &= A\hat{x}_1(t) + Bu(t) + \\ &\quad L\widetilde{\Theta}(t)\left[\alpha_1 C\widetilde{e}(t) - \alpha_2 Ce(t) - \alpha(\delta(t)+\nu(t))\right] \end{aligned} \tag{32}$$

also we rewrite (28) as

$$\dot{\hat{x}}_a(t) = A\hat{x}_a(t) + Bu(t) + LCs(t) + LC\widetilde{e}(t) - L(\delta(t)+\nu(t)) \tag{33}$$

then

$$\begin{aligned} \dot{s}(t) &= A\hat{x}_1(t) + Bu(t) + L\widetilde{\Theta}(t)\left[\alpha_1 C\widetilde{e}(t)\right. \\ &\quad \left. -\alpha_2 Ce(t) - \alpha\left(\delta(t)+\nu(t)\right)\right] - A\hat{x}_a(t) - \\ &\quad Bu(t) - LCs(t) - LC\widetilde{e}(t) - L(\delta(t)+\nu(t)) \\ &= (A-LC)s(t) + L(\alpha_1\widetilde{\Theta}(t) - I)C\widetilde{e}(t) + \\ &\quad \alpha_2 L\widetilde{\Theta}(t)Ce(t) + L(\alpha\widetilde{\Theta}(t) - I)(\delta(t)+\nu(t)). \end{aligned} \tag{34}$$

Without MTD, $E\left[\widetilde{\Theta}(t)\right] = I$, resulting in

$$\begin{aligned} \dot{s}(t) &= (A-LC)s(t) + L(\alpha_1 - 1)C\widetilde{e}(t) + \\ &\quad \alpha_2 LCe(t) + L(\alpha - 1)(\delta(t)+\nu(t)). \end{aligned} \tag{35}$$

For the first case of $\alpha_1 = 1$ and $\alpha_2 = 0$ the equation will be reduced to

$$\dot{s}(t) = (A-LC)s(t) \tag{36}$$

which is stable independent of $\delta(t)$ and $\lim_{t\to\infty} E\left[s(t)\right] \to 0$. Also, we have $E\left[\nu(t)\right] = 0$ which gives the corresponding detection statistics,

$$\lim_{t\to\infty} E\left[\widetilde{h}(t)\right] = \left|\bar{\widetilde{r}}(t)\right| = |\tau| \tag{37}$$

notice that this may never trigger the alarm, thus rendering the attack stealthy. The state's drift is limited by

$$\lim_{t\to\infty} E\left[x(t)\right] = \left(A+BK\right)^{-1} BK\left(A-LC\right) L\bar{\delta}(t) \tag{38}$$

where

$$\bar{\delta}(t) = -C(\bar{e}(t) + \bar{s}(t)) + \tau = -C\bar{e}(t) + \tau \tag{39}$$

and $\bar{e}(t)$ is from (17).

For the second case of $\alpha_1 = 1$ and $\alpha_2 = 1$ the equation will be reduced to

$$\dot{s}(t) = (A-LC)s(t) + \ LCe(t) + L(\delta(t)+\nu(t)) \tag{40}$$

which gives

$$\begin{aligned} \lim_{t\to\infty} \bar{s}(t) &= -\left(A-LC\right)^{-1}\left[LC\bar{e}(t) + L\bar{\delta}(t)\right] \\ &= -\left(A-LC\right)^{-1}\left[LC\bar{e}(t) - LC\bar{e}(t) + L\tau\right] \\ &= -\left(A-LC\right)^{-1} L\tau \end{aligned} \tag{41}$$

this is a constant and does not change with time.

$$\begin{aligned} \lim_{t\to\infty} E\left[\widetilde{h}(t)\right] &= \left|\bar{\widetilde{r}}(t)\right| \\ &= \left|-C\bar{s}(t) + \tau\right| \\ &= \left|C\left(A-LC\right)^{-1} L\tau + \tau\right| \end{aligned} \tag{42}$$

For the case with MTD, we will apply the expectation operator on (34)

$$\begin{aligned} \dot{\bar{s}}(t) &= (A-LC)\bar{s}(t) + L(\alpha_1\widetilde{P} - I)C\bar{\widetilde{e}}(t) + \\ &\quad \alpha_2 L\widetilde{P}C\bar{e}(t) + L(\alpha\widetilde{P} - I)\bar{\delta}(t) \\ &= (A-LC)\bar{s}(t) + L(\alpha_1\widetilde{P} - I)C\bar{\widetilde{e}}(t) + \\ &\quad (\alpha_2 L\widetilde{P} - I)C\bar{e}(t) + L(\alpha\widetilde{P} - I)\tau. \end{aligned} \tag{43}$$

$$\begin{aligned} \lim_{t\to\infty} \bar{s}(t) &= -\left(A-LC\right)^{-1}\left[L(\alpha_1\widetilde{P} - I)C\bar{\widetilde{e}}(t) + \right. \\ &\quad \left. (\alpha_2 L\widetilde{P} - I)C\bar{e}(t) + L(\alpha\widetilde{P} - I)\tau\right] \end{aligned} \tag{44}$$

and $\bar{\widetilde{e}}(t)$ will be

$$\lim_{t\to\infty} E\left[\widetilde{e}(t)\right] = \left(A - \alpha_1 L\widetilde{P}C\right)^{-1} L\widetilde{P}\left(\alpha\bar{\delta}(t) + \alpha_2 C\bar{e}(t)\right). \tag{45}$$

Taking the expectation of (31) and substitute

$$\begin{aligned} \bar{\widetilde{h}}(t) &= \lim_{t\to\infty} |\bar{\widetilde{r}}(t)| \\ &= \left|C\left(A-LC\right)^{-1}\left[L(\alpha_1\widetilde{P} - I)C\bar{\widetilde{e}}(t) + \right.\right. \\ &\quad \left.\left. (\alpha_2 L\widetilde{P} - I)C\bar{e}(t) + L(\alpha\widetilde{P} - I)\tau\right] + \tau\right| \end{aligned} \tag{46}$$

first for case of $\alpha_1 = 1$ and $\alpha_2 = 0$

$$\begin{aligned} \bar{\widetilde{h}}(t) &= \lim_{t\to\infty} |\bar{\widetilde{r}}(t)| \\ &= \left|C\left(A-LC\right)^{-1}\left[L(\widetilde{P} - I)C\bar{\widetilde{e}}(t) + \right.\right. \\ &\quad \left.\left. L(\widetilde{P} - I)\tau\right] + \tau\right| \end{aligned} \tag{47}$$

Recall the state's dynamics from (12) to be

$$\dot{x}(t) = (A + BK)x(t) - BKe(t) + w(t) \tag{48}$$

and the expected value for the error $e(t)$ will be

$$\lim_{t\to\infty} \bar{e}(t) = (A - LPC)^{-1} LP\bar{\delta}(t) \tag{49}$$

which gives the following equilibrium expression for the states after substituting $\bar{e}(t)$ from (49)

$$\lim_{t\to\infty} E\left[x(t)\right] = (A + BK)^{-1} BK\left(A - LPC\right) LP\bar{\delta}(t) \tag{50}$$

where $\bar{\delta}(t)$ from (39).

Now, for case of $\alpha_1 = 1$ and $\alpha_2 = 1$ we will have

$$\begin{aligned} \bar{\tilde{h}}(t) \quad = \quad & \Big| C\left(A - LC\right)^{-1} \Big[L(\widetilde{P} - I)C\bar{\tilde{e}}(t) + \\ & (L\widetilde{P} - I)C\bar{e}(t) + 2L(\widetilde{P} - I)\tau \Big] + \tau \Big| \end{aligned} \tag{51}$$

and (50) will be the same.

IV. MTD DESIGN

The design revolves around executing a sequence of two optimization processes to enhance and refine the overall performance. We used a simple grid search to find the optimal values of each optimization, while [7] suggested using the interior-point algorithm. However, proposing the most suitable optimization algorithm is an interesting research direction but lies beyond this paper's scope and will not be considered here. We also assume that the values of the parameters α_1 and α_2 are already set by the designer based on experience or trial-and-error; they are not optimized in the following sections.

A. Optimize attack impact on states

As a first step, we determine the optimal parameter P^* for the first MTD. This parameter aims to minimize attacks' impact on states while guaranteeing the system's stability. We are optimizing here the impact of attacks on states in case of fixed attack (27), given the dynamic system parameters, controller matrix K, observer gain L, and the detection threshold vector τ, i.e.

$$\begin{aligned} P^* = & \max_{P} \quad \|\bar{x}(t)\|^{-1} \equiv \\ = & \max_{P} \quad \| \left(A + BK\right)^{-1} BK\left(A - LPC\right) LP\tau\|^{-1} \\ & \text{s.t.} \quad \Re\{\lambda_{\max}(A - LPC)\} \leq \Re\{\lambda_{\max}(A - LC)b\} < 0 \\ & 0 < p_i \leq 1, \quad \forall i \in S \end{aligned} \tag{52}$$

where $\Re\{\}$ is the real part of a complex number, and $b < 1$ is a design parameter such that $\Re\{\lambda_{\max}(A - LC)\}b$ is an upper bound for the largest negative eigenvalue of the new system that determines the slowdown in convergence speed due to MTD mechanism.

B. Optimize for attack detection

In the second step, we find the parameters $\widetilde{P}^*$ for the second MTD that achieve optimal attack detection while maintaining system stability,

$$\begin{aligned} \widetilde{P}^* = & \max_{\widetilde{P}} \quad \|\bar{\tilde{h}}(t)\| \\ & \text{s.t.} \quad \Re\{\lambda_{\max}(\bar{F})\} \leq \Re\{\lambda_{\max}(A - LC)b\} < 0 \\ & 0 < \widetilde{p}_i \leq 1, \quad \forall i \in S \end{aligned} \tag{53}$$

where$\bar{F}$ from (13) is

$$\bar{F} := \begin{bmatrix} A - LPC & 0 \\ \alpha_2 L\widetilde{P}C & A - \alpha_1 L\widetilde{P}C \end{bmatrix}. \tag{54}$$

The optimization is restricted by maintaining identical upper bounds on eigenvalues to prevent slowing down the convergence speed. To ensure stealth attack detection, we need to use the stealth attack expression (46) for $\bar{\tilde{h}}(t)$.

For both optimizations, knowing the threshold vector τ is required. This may be calculated from noise statistics at system outputs and a given P_{FA}. For Gaussian noise distribution, the expression is not tractable, and approximation is required, similar to (10).

V. CASE STUDIES

We consider here the following linear time-invariant system, which is the same example in [7] to compare easily

$$\begin{aligned} A \quad = \quad & \begin{bmatrix} 1 & 0.5 & 0.4 \\ 0.3 & -2 & -0.5 \\ 0.1 & 1 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \\ C \quad = \quad & \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \end{aligned} \tag{55}$$

with unitary costs LQR controller

$$K = \begin{bmatrix} -6.2 & -1.23 & -0.77 \\ -4 & -0.88 & -0.35 \end{bmatrix} \tag{56}$$

and Kalman filter gain

$$L = \begin{bmatrix} 2.0726 & 0.3431 \\ 0.2040 & 0.5216 \\ 0.1312 & 0.1362 \end{bmatrix}. \tag{57}$$

We added zero mean Normally distributed noise to each output with a variance of 10^{-6}. The matrix A has three eigenvalues, one positive and one negative complex conjugate pair, which makes the original system unstable. To clarify the advantages of the presented MTD, we shall divide the simulation results into two subsections, one for the fixed attack and one for the stealth attack.

A. Fixed attack performance

The main results of this paper are summarized in Figure 2. The upper plot shows the attack impact on states, and the lower plot shows the detection statistics for the bad-data detector. It is easy to see the improvement in the attack impact compared to the case of no MTD and also to the original MTD presented by [7]. At the same time, the bottom plot shows improvement in detection statistics

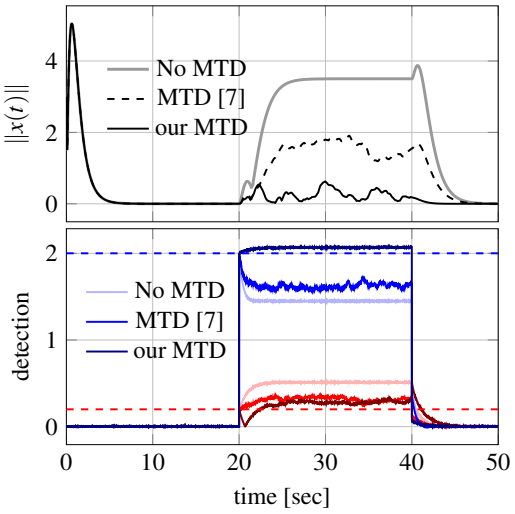

Fig. 2. Our proposed MTD improves attack impact on states and detection statistics at the same time. The red and blue curves are for h_1 and h_2, respectively, while the dashed lines are for the detection thresholds. The attack is initiated at time 20 sec and stopped at 40 sec.

compared to others, as we have larger margins than the detection threshold in the dashed lines. Notice that the price for this improvement in the detection is a slower detection rise time. Notice also that detection statistics for the second output (red curves) are reduced with both MTD algorithms compared to the case without MTD, if this is not desirable, it can be controlled by replacing the second norm with the weighted norm in the optimization. We used to generate those plots the optimized values of $P^* = \text{diag}([0.81, 0.35])$, and $\widetilde{P}^* = \text{diag}([0.88, 0.0])$ which obtained for $b = 0.9$ and $\tau = \begin{bmatrix}-0.2 & 2.0\end{bmatrix}^{\mathrm{T}}$. The above results were obtained for $\alpha_1 = 1$ and $\alpha_2 = 0$.

We repeat the same simulations but for different observer parameters $\alpha_1 = 1$ and $\alpha_2 = 1$. Figure 3 presents detection thresholds comparison between the proposed MTD and without MTD case. It is clear that the second output statistic is not reduced.

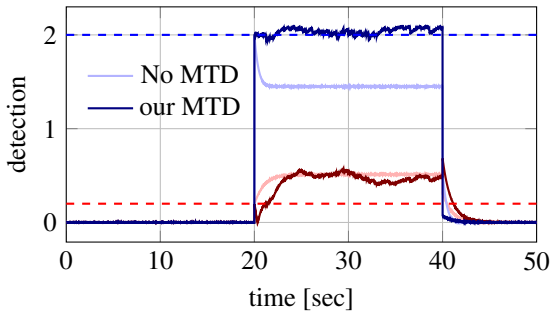

Fig. 3. Detection statistics for $\alpha_1 = 1$ and $\alpha_2 = 1$. We used $P^* = \text{diag}([0.81, 0.35])$, and $\widetilde{P}^* = \text{diag}([1.0, 0.04])$.

We used the same above parameters for simulations with $\alpha_1 = 3$ and $\alpha_2 = 3$, Figure 4 presents the case with MTD and without MTD. Now, both statistics are slightly improved.

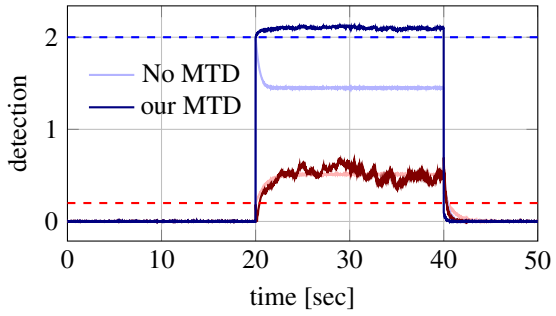

Fig. 4. Detection statistics for $\alpha_1 = 3$ and $\alpha_2 = 3$. We used $P^* = \text{diag}([0.81, 0.35])$, and $\widetilde{P}^* = \text{diag}([0.52, 0.0])$.

Therefore, changing the parameters α_1 and α_2 has a large influence on the resulting performance. Proper selection of those parameters is important to improve the performance further.

B. Stealth attack performance

We repeat the above cases with the same system and attack parameters but for the stealth attack case presented in Section III-C. Figures 5, 6, 7 show the obtained results where our proposed MTD pass the threshold for both statistics in the three figures while only one passes the threshold for the MTD in [7]. Notice that the stealth attack causes the state norm to increase indefinitely after a few seconds from the attack in Figure 5. This is because the original matrix A has a positive eigenvalue that makes the feedback system unstable as explained in Section III-C.

VI. Conclusions

This paper introduces a novel structure for a switching-based MTD system designed to enhance security. Our system significantly complicates an attacker's ability to predict the controller output and the outcome of attack detection. We have demonstrated the effectiveness of our approach in identifying highly stealthy attacks, even when attackers possess complete knowledge of system dynamics and detection strategies. Our approach leverages two MTD algorithms: one introduces uncertainty into the controller to mitigate the impact of false-data-injection attacks on system states, while the other introduces uncertainty into the IDS to enhance attack detection. We formulated two optimization problems to address these challenges, one for each MTD algorithm.

Through a case study of a linear time-invariant system, we have assessed and showcased the merits of our proposed algorithm.

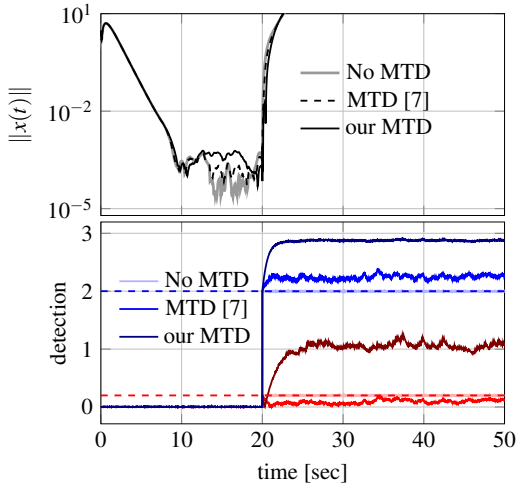

Fig. 5. Our proposed MTD improves detection statistics in case of stealth attack. The red and blue curves are for h_1 and h_2, respectively, while the dashed lines are for the detection threshold. The attack was initiated at a time of 20 seconds.

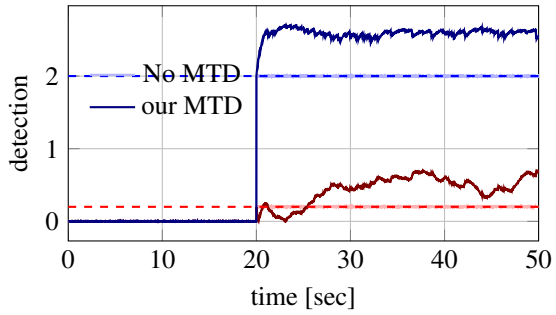

Fig. 6. Detection statistics for $\alpha_1 = 1$ and $\alpha_2 = 1$. We used $P^* = \text{diag}([0.81, 0.35])$, and $\widetilde{P}^* = \text{diag}([1.0, 0.04])$.

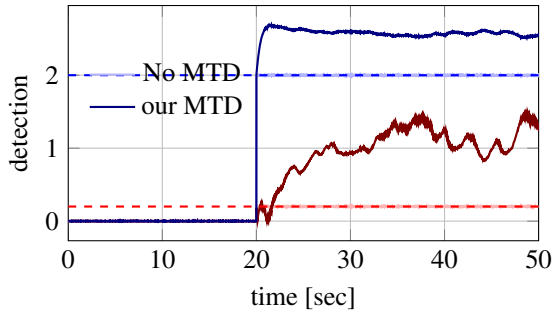

Fig. 7. Detection statistics for $\alpha_1 = 3$ and $\alpha_2 = 3$. We used $P^* = \text{diag}([0.81, 0.35])$, and $\widetilde{P}^* = \text{diag}([0.52, 0.0])$.

REFERENCES

[1] J. Slay and M. Miller, "Lessons learned from the maroochy water breach," in *International conference on critical infrastructure protection*. Springer, 2007, pp. 73–82.

[2] J. P. Farwell and R. Rohozinski, "Stuxnet and the future of cyber war," *Survival*, vol. 53, no. 1, pp. 23–40, 2011.

[3] S. Y. A. Fayi, "What petya/notpetya ransomware is and what its remidiations are," in *Information Technology-New Generations: 15th International Conference on Information Technology*. Springer, 2018, pp. 93–100.

[4] A. Di Pinto, Y. Dragoni, and A. Carcano, "Triton: The first ics cyber attack on safety instrument systems," *Proc. Black Hat USA*, vol. 2018, pp. 1–26, 2018.

[5] R. Zhuang, S. A. DeLoach, and X. Ou, "Towards a theory of moving target defense," in *Proceedings of the first ACM workshop on moving target defense*, 2014, pp. 31–40.

[6] J. Giraldo, A. Cardenas, and R. Sanfelice, "A moving target defense to reveal cyber-attacks in cps and minimize their impact," in *American Control Conference*, 2019.

[7] ——, "A switching-based moving target defense against sensor attacks in control systems," *Nonlinear Analysis: Hybrid Systems*, vol. 47, p. 101268, 2023.

[8] Y. Wan and J. Cao, "A brief survey of recent advances and methodologies for the security control of complex cyber–physical networks," *Sensors*, vol. 23, no. 8, p. 4013, 2023.

[9] J. Zheng and A. S. Namin, "A survey on the moving target defense strategies: An architectural perspective," *Journal of Computer Science and Technology*, vol. 34, pp. 207–233, 2019.

[10] M. Zhang, X. Fan, R. Lu, C. Shen, and X. Guan, "Extended moving target defense for ac state estimation in smart grids," *IEEE Transactions on Smart Grid*, vol. 14, no. 3, pp. 2313–2325, 2022.

[11] M. A. Rahman, E. Al-Shaer, and R. B. Bobba, "Moving target defense for hardening the security of the power system state estimation," in *Proceedings of the First ACM Workshop on Moving Target Defense*, 2014, pp. 59–68.

[12] S. Weerakkody and B. Sinopoli, "Detecting integrity attacks on control systems using a moving target approach," in *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE, 2015, pp. 5820–5826.

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

[14] B. Satchidanandan and P. R. Kumar, "Dynamic watermarking: Active defense of networked cyber–physical systems," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 219–240, 2017.

[15] H. Liu, Y. Zhang, Y. Li, and B. Niu, "Proactive attack detection scheme based on watermarking and moving target defense," *Automatica*, vol. 155, p. 111163, 2023.

[16] A. Kanellopoulos and K. G. Vamvoudakis, "A moving target defense control framework for cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 65, no. 3, pp. 1029–1043, 2019.

[17] M. Liu, C. Zhao, Z. Zhang, R. Deng, P. Cheng, and J. Chen, "Converter-based moving target defense against deception attacks in dc microgrids," *IEEE Transactions on Smart Grid*, vol. 13, no. 5, pp. 3984–3996, 2021.

[18] J. A. Giraldo, M. El Hariri, and M. Parvania, "Moving target defense for cyber–physical systems using iot-enabled data replication," *IEEE Internet of Things Journal*, vol. 9, no. 15, pp. 13 223–13 232, 2022.

[19] J. Giraldo, M. El Hariri, and M. Parvania, "Decentralized moving target defense for microgrid protection against false-data injection attacks," *IEEE Transactions on Smart Grid*, vol. 13, no. 5, pp. 3700–3710, 2022.

[20] Y. Zhao and F. Zhu, "Security control of cyber-physical systems under denial-of-service sensor attack: A switching approach," in *2021 IEEE 10th Data Driven Control and Learning Systems Conference (DDCLS)*. IEEE, 2021, pp. 1112–1117.

[21] K. G. Vamvoudakis, J. P. Hespanha, B. Sinopoli, and Y. Mo, "Detection in adversarial environments," *IEEE Transactions on Automatic Control*, vol. 59, no. 12, pp. 3209–3223, 2014.

[22] H. Liu, S. Wang, and Y. Li, "Event-triggered control and proactive defense for cyber–physical systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 10, pp. 6305–6313, 2022.

[23] C. Peng and H. Sun, "Switching-like event-triggered control for networked control systems under malicious denial of service attacks," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3943–3949, 2020.

[24] D. Chatterjee and D. Liberzon, "Stabilizing randomly switched systems," *SIAM Journal on Control and Optimization*, vol. 49, no. 5, pp. 2008–2031, 2011.