A Switching-based Moving Target Defense Against Sensor Attacks in Control Systems

J. Giraldo[a,*], A. Cardenas[b,1], R.G. Sanfelice[b,1]

[a] *University of Utah, 50 S Central Campus Dr, Salt Lake City, UT, 84112*
[b] *University of California Santa Cruz, 1156 High Street, Santa Cruz, CA 95064*

Abstract

Moving Target Defense (MTD) prevents adversaries from being able to predict the effect of their attacks by adding uncertainty in the state of a system during runtime. In this paper, we present an MTD algorithm that randomly changes the availability of the sensor data, so that it is difficult for adversaries to tailor stealthy attacks while, at the same time, minimizing the impact of false-data injection attacks. Using tools from the design of state estimators, namely, observers, and switched systems, we formulate an optimization problem to find the probability of the switching signals that increase the visibility of stealthy attacks while decreasing the deviation caused by false data injection attacks. We show that the proposed MTD algorithm can be designed to guarantee the stability of the closed-loop system with desired performance. In addition, we formulate an optimization problem for the design of the parameters so as to minimize the impact of the attacks. The results are illustrated in two case studies, one about a generic linear time-invariant system and another about a vehicular platooning problem.

Keywords: Cyber-security, Moving Target Defense, Industrial Control Systems, Sensor Attacks

1. Introduction

1.1. Motivation and Related Work

The emergence of physical systems that are controlled and coordinated over networks by computer algorithms has propelled the development of security tools that can determine if both physical and cyber components are shielded from attacks. Such systems sometimes referred to as *computer-controlled systems*, more recently called *cyber-physical systems* (CPSs), require algorithms

*Corresponding author

that operate in contested environments, where attackers may attempt to compromise their operation.

While several frameworks for analysis and design of secure control systems have recently been developed, including the work in [1, 2] using a static representation of systems, and dynamic approaches in [3, 4] using game theoretical tools, in [5, 6] using information theory, in [7, 8] using classical control theoretical tools for discrete-time systems or continuous-time systems, and in [9, 10] using switched/hybrid systems tools, these systems do not provide defense-in-depth against a knowledgeable and strategic adversary. This powerful adversary knows the details of the control algorithm, the model of the plant, and the defense mechanism proposed. Therefore, if the attacker can predict the behavior of the system under attack, the success of the attack will be more likely.

To overcome this vulnerability, Moving Target Defense (MTD) has emerged as a strategy to add uncertainty about the state and execution of a system in order to prevent adversaries from having predictable effects with their attacks and also to improve the chances of detecting stealthy attacks [11]. According to the National Science and Technology Council, MTD

> *"enables us to create, analyze, evaluate, and deploy mechanisms and strategies that are diverse and that continually shift and change over time to increase complexity and cost for attackers, limit the exposure of vulnerabilities and opportunities for attack, and increase system resiliency. The characteristics of an MTD system are dynamically altered in ways that are manageable by the defender yet make the attack space appear unpredictable to the attacker." [12]*

Several authors have proposed MTD approaches for state estimation in the smart grid [13, 14, 15], where the main idea consists of changing the physical topology of the power grid in order to reveal false data injection attacks. A similar idea uses an authenticating signal in the control of the system, and if an anomaly detection system does not detect this signal watermark in the sensor readings, then it raises an alert [16, 17]. Finally, another set of MTD strategies increase the uncertainty of the system by randomly switching among several controllers [18], varying the power system configuration [19], adding random noise to the controller in order to make harder for an adversary to estimate the controller output [20], or using IoT devices to replicate sensor data that is randomly transmitted [21].

1.2. Contribution

In this paper, we propose the use of MTD for a class of cyber-physical systems. By combining ideas from state estimation for linear systems and stability results for switched systems, we generate techniques and implementable algorithms to detect attacks and minimize the impact on the system. Specifically, in this paper, we propose an MTD algorithm with the following properties:

- *Detection of attacks with high accuracy.* Starting from the premise that it should be hard for the attacker to evade an intrusion detection system

(IDS), we propose a methodology that utilizes an MTD algorithm in order to facilitate the detection of strong stealthy attacks [22, 23] . Our method works even when the adversary knows the system dynamics, the attack-detection strategy, and has access to all control inputs, and all sensor readings.

- *Minimization of the impact of a sensor compromise.* Even when the attacker compromises a sensor in a control system, we show that once the proposed MTD algorithm is activated, the impact of the attack can be minimized in a controlled manner. For this purpose, we formulate an optimization problem for the design of the parameters in our MTD algorithm with the objective of minimizing the impact of attacks.

One of the key features of our proposed method is that it can be designed in a way that stability of the original control system, namely, the attack-free system without an MTD strategy, is preserved and performance of the resulting system with our MTD algorithm is specified by design. For this purpose, we formulate conditions guaranteeing asymptotic stability and an optimization problem that incorporates information about system performance. Another salient feature of our MTD algorithm is that it does not have to be active at all times. In fact, our MTD algorithm might be activated only when external indicators suggest the presence of an attack.

A preliminary version of this work appeared in the conference publication [24], and we have extended it by including additional insights on the derivation of the stability condition and the design optimization problem. We have introduced an additional simplified optimization problem that it is easy to solve but it is applicable only for a low number of sensors. Furthermore, the proposed MTD strategy is also applied to a vehicular platooning scenario, and we show that the proposed MTD can help to avoid crashes.

1.3. Organization and Notation

The class of systems and attack models, as well as the proposed MTD strategy, are presented in Section 2. The conditions for stability of the resulting system with MTD are given in Section 3. Section 4 shows how the proposed MTD algorithm enables the detection of stealthy attacks. Design conditions for the MTD algorithm are given in Section 5, where, in particular, we present an optimization problem for the minimization of the impact of attacks.

Notation: The n-th dimensional Euclidean space is denoted $\mathbb{R}^n$. The diagonal matrix with diagonal entries $\theta_1, \theta_2, \ldots, \theta_q$ is denoted $\text{diag}(\theta_1, \theta_2, \ldots, \theta_q)$. The operator $E[\cdot]$ denotes the expectation operator that takes a random variable and returns its average value. For a matrix $A \in \mathbb{R}^{n\times n}$, let $\lambda_1, \ldots, \lambda_n$ be their respective eigenvalues. Then, $\lambda_{max} = \max_{i=1,\ldots,n}(Re(\lambda_i))$, for $Re(\cdot)$ the real part of a complex number corresponds. Similarly, $\lambda_{min} = \min_{i=1,\ldots,n}(\lambda_i)$. The set of positive integers including zero is denoted $\mathbb{Z}_+$. The norm operator $\|\cdot\|$ refers to the Euclidean norm. A class $\mathcal{K}$ function $\alpha : [0, a] \to [0, \infty]$ is strictly increasing and $\alpha(0) = 0$. If $a = \infty$, and $\alpha(r) \to \infty$ as $r \to \infty$, then α is said to be a $\mathcal{K}_\infty$ function.

2. Modeling and Proposed MTD Algorithm

We consider the control system depicted in Fig.1 which consists of a physical process (i.e., plant) that possesses sensors and actuators, a moving target defense (MTD) mechanism that randomizes which sensor values the controller uses at a given time, an observer-based controller that uses the available sensor measurements modified by the MTD mechanism to compute the estimation of the system states and calculate control commands, and an intrusion detection system (IDS). The main goal of the MTD mechanism is to add uncertainty to the system so it is harder for the attacker to hide its attacks and simultaneously limit the impact of the attack; namely, to limit how much control the adversary gets over the plant.

2.1. System Description

We consider a continuous-time linear time-invariant systems of the form

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

where $t \mapsto x(t) \in \mathbb{R}^n$, $t \mapsto u(t) \in \mathbb{R}^m$, $t \mapsto y(t) \in \mathbb{R}^q$ are the states, input, and output vectors, respectively. The signal $t \mapsto \delta^a(t) \in \mathbb{R}^q$ denotes the attack vector injected to the sensors. The signal $t \mapsto \widetilde{y}(t) \in \mathbb{R}^q$ is the output received by the estimator, where $t \mapsto \Theta(t)$ denotes the MTD mechanism.

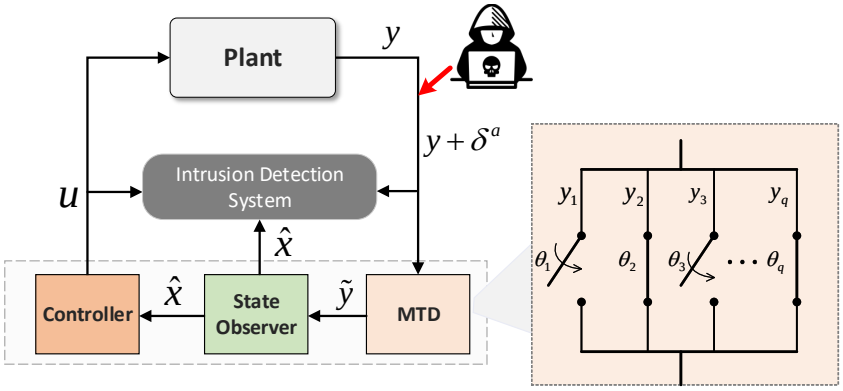

Figure 1: General architecture of the feedback control system with the proposed MTD strategy.

2.2. Moving Target Defense Mechanism

We propose an MTD approach that randomly changes the availability of the sensors as depicted in Fig. 1. Let $t \mapsto \Theta(t)$ be a diagonal matrix that, for each $t \geq 0$, is of the form $\Theta(t) = \text{diag}(\theta_1(t), \theta_2(t), \ldots, \theta_q(t))$ and let $\mathcal{S} = \{1, 2, \ldots, q\}$

be the sensor index set, where q is an integer larger than or equal to one. Therefore,

$$\begin{aligned}\widetilde{y}_i(t) &= \theta_i(t) y_i(t) \\ &= \theta_i(t)(C_i x(t) + \delta_i(t)),\end{aligned} \tag{2}$$

for all $i \in \mathcal{S}$, where $C_i \in \mathbb{R}^{1\times n}$ denotes the i^{th} row of matrix C and, for each $t \geq 0$, $\theta_i(t) \in \{0, 1\}$ is a piecewise binary signal. In particular, we focus our attention on *random switching*, where we only need to define the probability distribution of a group of Bernoulli random variables.

Let $\mathcal{T} = \{t_0, t_1, \ldots, t_k, \ldots\}$ denote the set of time instances at which the MTD strategy is updated, satisfying $0 < T_{min} < t_k - t_{k-1} < T_{max}$, where T_{min} and T_{max}. Moreover, let $\{S_k\}_{k\in\mathbb{Z}_+}$ be the sequence of holding times where $S_{k+1} = t_{k+1} - t_k$. We can define $\beta_j(t_k) \sim \mathcal{B}(p_j)$ as a random variable drawn from a Bernoulli distribution such that $\beta_j(t_k) = 1$ with probability p_j (and zero otherwise) for all $t_k \in \mathcal{T}$. Therefore, we have that on each time interval $[t_k, t_{k+1})$, $k \in \mathbb{Z}_+$,

$$\theta_j(t) = \beta_j(t_k) \;\; \forall t \in [t_k, t_{k+1}).$$

Remark 2.1. *We will see later that the sequence $\{S_k\}_{k\in\mathbb{Z}_+} = t_{k+1} - t_k$ is considered random, which adds an extra level of uncertainty to the MTD strategy, making it even harder for an adversary to predict the system's behavior.*

2.3. State-Observer and Control with MTD Measurements

We propose a state observer given by

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L\Theta(t)(\widetilde{y}(t) - C\hat{x}(t)), \tag{3}$$

which measures $t \mapsto \widetilde{y}(t)$, which switches over time, as well as the input $t \mapsto u(t)$ and the state $t \mapsto \hat{x}(t)$. When the pair (C, A) is detectable and the function $t \mapsto \Theta(t)$ is properly designed (the conditions for the design of Θ will be introduced in Section 3), the gain $L \in \mathbb{R}^{n\times q}$ can be designed to reconstruct the system state. Since the values that Θ assumes at each t are under full control of our algorithm, the observer in (3) is aware of which sensor readings are active – through knowledge of $\Theta(t)$ – and updates its estimation using only those active readings. Let us define, for each $t \geq 0$, $e(t) = x(t) - \hat{x}(t)$ as the estimation error.

Combining (1) and (3), and since $\Theta(t)\Theta(t) = \Theta(t)$ we obtain

$$\dot{e}(t) = (A - L\Theta(t)C)e(t) - L\Theta(t)\delta^a(t), \tag{4}$$

and the observer design becomes a stabilization problem where L and $\Theta(t)$ have to be chosen in such a way that the switched system in (4) has $e = 0$ globally asymptotically stable when $\delta^a(t) = 0$ for all $t \geq 0$. Finally, we consider that the pair (A, B) is controllable, and the output-feedback controller of the form

$$u(t) = K\hat{x}(t), \tag{5}$$

for each $t \geq 0$.

Remark 2.2. *In this work we assume that the gain L is given and fixed in order to consider the worst case scenario for the defender, such that the attacker also knows L. In addition, it is desired to design the proposed MTD with existing system parameters to facilitate its seamless integration with the system.*

2.4. Intrusion Detection

Taking advantage of the state estimator in (3), we can construct anomaly detection modules that compare the estimated sensor readings with the real ones to determine the presence of an attack. Therefore, we define the residuals as

$$\begin{aligned} r(t) &= y(t) - C\hat{x}(t) \\ &= Ce(t) + \delta^a(t). \end{aligned} \tag{6}$$

The anomaly detection then takes the vector of residuals $r(t)$ and computes a measure of how deviated the sensor readings are from the estimation. There are different types of anomaly detection strategies, such as the χ^2-test, distributed bad-data detection, and CUSUM [23]. For simplicity, we will focus our attention on the distributed bad-data detection with the detection statistic given by

$$h(t) = |r(t)|, \tag{7}$$

where $|\cdot|$ is evaluated component-wise. If any $h_i(t) > \tau_i$, for some fixed detection threshold $\tau_i > 0$, then an alarm is triggered. Particularly, each τ_i is selected to maintain a tolerable false-alarm rate under normal operation. Note that, since the proposed MTD affects state estimation, τ_i would be chosen to be larger than for the case without MTD.

Remark 2.3. *We have omitted the effect of system and sensor noise in our formulation to focus on how the proposed MTD approach provides a nominal defense mechanism against cyberattacks. Adding noise would only affect the selection of τ_i but it would not affect the proposed MTD design.*

2.5. Adversary Model

In this work we consider a motivated and resourceful adversary that intents to disrupt the normal system operation. The adversary's capabilities and knowledge are as follows.

Capabilities and goals: the attacker has gained access to a set of sensors and is able to inject false signals δ^a to drive the system away from the operational states. This can be achieved by introducing malware in monitoring devices or performing man-in-the-middle-attacks in the communication network between the sensor and the controller. We assume the adversary has reading access to the control commands $u(t)$ in order to construct sophisticated stealthy attacks.

Knowledge: the attacker knows the non-MTD system model; i.e., the attacker knows A, B, C, K, the estimation gain L, and the detection threshold τ

but does not know the existence of the MTD mechanism (i.e., the attacker does not know Θ). Also, the attacker does not have access to the state estimate $\hat{x}$ used for the controller and for the IDS. These strong assumptions allow us to consider worst-case scenarios, implying that our MTD will be effective against weaker adversaries.

2.6. Motivational Example

To illustrate why our MTD approach can make it difficult for an adversary to design strong attacks and also minimize the impact of an attack in the system states, we consider the simple example where $a = -0.1, b = 1, L = 0.2, K = -0.3, C = 1$,with an intrusion detection threshold of $\tau_i = \tau = 0.1$.

Without MTD. $\Theta(t) = 1$, and if $\delta^a = 0.3$, then the detection statistic in the limit converges to

$$\lim_{t\to\infty} h(t) = |c(a - Lc)^{-1}L\delta^a + \delta^a| = 0.1,$$

and therefore the attack remains stealthy (undetected by our residual-based IDS). We can measure the impact of the attack in terms of how much the system state deviates from the origin. Without MTD, we have that

$$\lim_{t\to\infty} x(t) = (a + bK)^{-1}bK(a - Lc)^{-1}L\delta^a = -0.15$$

With MTD. With the proposed MTD mechanism where $\Theta(t) = \theta(t)$ with $p = 0.3$, and $\delta^a = 0.3$, we have that

$$\lim_{t\to\infty} E[h(t)] = |c(a - Lpc)^{-1}Lp\delta^a + \delta^a| = 0.1875,$$

and *the same attack is no longer stealthy.*

Now, with the MTD mechanism, the expected state converges to

$$\lim_{t\to\infty} E[x(t)] = (a + bK)^{-1}bK(a - LpC)^{-1}Lp\delta^a = -0.084.$$

Note that for this particular example the random MTD mechanism causes the residuals to increase while the state deviation decreases which illustrates the potential benefits of the proposed approach.

The cost of MTD can be observed in terms of the convergence speed. In our example, the slowest (and only) eigenvalue of the expected estimation error is $\lambda_{MTD} = a - LpC$ and without MTD is $\lambda_{noMTD} = a - LC$. Clearly $\lambda_{MTD} > \lambda_{noMTD}$ for any $0 \leq p < 1$ and therefore the observer convergence is degraded when p is small.

In the next section, we formulate our problem as a switched system and derive conditions for stability.

3. Stability of the MTD System

3.1. Switched System

In order to formulate our problem as a switched system and exploit some existing tools, we define the family of non-identical diagonal binary matrices $\{\Theta_1, \Theta_2, \ldots, \Theta_s\}$, and the finite index set $\Sigma = \{1, 2, \ldots, s\}$, where $s = 2^q$. Each $\Theta_i \in \mathbb{R}^{q\times q}$ describes one possible combination of $\{1, 0\}$ for each $\theta_1, \theta_2, \ldots, \theta_q$, where $i \in \Sigma$. We also define the piecewise switching signal $\sigma : [0, \infty) \to \Sigma$, which is updated at the time points $t_k \in \mathcal{T}$ and remains constant in the time interval (t_k, t_{k+1}). The signal $t \mapsto \sigma(t)$ is used to specify, at each time instant t, the index $i \in \Sigma$ of each active subsystem.

Then, our MTD approach in (2) can be rewritten as

$$\widetilde{y}(t) = \Theta_{\sigma(t)} y(t), \tag{8}$$

where $\sigma(t)$ randomly chooses among the index set Σ, according to the probability mass function $\Omega : \Sigma \to [0, 1]$, where, for each $i \in \Sigma$,

$$\begin{aligned}\Omega(i) = \widetilde{p}_i &= \prod\nolimits_{j\in\mathcal{S}} [\Theta_i]_j p_j + (1 - [\Theta_i]_j)(1 - p_j) \\ &= \prod\nolimits_{j\in\mathcal{S}} (1 - p_j - [\Theta_i]_j + 2[\Theta_i]_j p_j),\end{aligned} \tag{9}$$

for $[\Theta_i]_j$ refers to the j^{th} diagonal element of matrix Θ_i.

Example: Let $p_i = p$, and the number of sensors is $q = 2$. Then there exist 4 possible matrices Θ_i, given by $\Theta_1 = \text{diag}(0, 0)$, $\Theta_2 = \text{diag}(1, 0)$, $\Theta_3 = \text{diag}(0, 1)$, $\Theta_4 = \text{diag}(1, 1)$, with a probability mass function $\Omega(i) = \{(1-p)^2, p(1-p), p(1-p), p^2\}$. for all $i \in \Sigma$

Having formulated our MTD strategy as a switched system, we can rewrite the observer in (3) as follows

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L\Theta_{\sigma(t)}(\widetilde{y}(t) - C\hat{x}(t)), \tag{10}$$

and the estimation error can be described by

$$\dot{e} = (A - L\Theta_{\sigma(t)}C)e - L\Theta_{\sigma(t)}\delta^a(t). \tag{11}$$

Let us define $F_{E,\sigma(t)} = A - L\Theta_{\sigma(t)}C$, and let $z(t) = [x^\top(t), e^\top(t)]^\top$ denote the extended state vector, such that

$$\begin{aligned}\dot{z} &= \begin{bmatrix} A + BK & -BK \\ 0 & F_{E,\sigma(t)} \end{bmatrix} z + \begin{bmatrix} 0 \\ -L\Theta_{\sigma(t)} \end{bmatrix} \delta^a \\ &=: F_{\sigma(t)} z + G_{\sigma(t)}\delta^a.\end{aligned} \tag{12}$$

Thanks to the separation principle, we can design K independently of the observer gain or the switching signal (e.g., an LQR that satisfies specific performance conditions). Therefore, if K is such that $A + BK$ is stable, the stability of (12) is dictated by $F_{E,\sigma(t)}$.

3.2. Stability Conditions

Assume $\delta^a(t) = 0$ for $t \geq 0$. Recall that we have the family of matrices $F_{E,i} = A - L\Theta_i C$ for all $i \in \Sigma$. The following lemma describes sufficient conditions to guarantee global uniform asymptotic stability (GUAS) [25].

Lemma 3.1. *Suppose that each $F_{E,i}$ is Hurwitz stable for all $i \in \Sigma$. If there exists $Q = Q^\top > 0$ such that*

$$F_{E,i}^\top Q + Q F_{E,i} < 0, \quad \textit{for all } i \in \Sigma,$$

then, there exists a quadratic common Lyapunov function $e \mapsto V(e)$ and the equilibrium $e = 0$ is GUAS for any arbitrary switching.

From Lemma 3.1, stability is guaranteed if we can find an adequate gain L. However, the limitation of the conditions stated in Lemma 3.1 lies in the fact that it requires all subsystems represented by $F_{E,i}$ are Hurwitz stable, which may not be feasible if the pair $(A, \Theta_i C)$ is not observable for any $i \in \Sigma$. In addition, for the particular case when $\Theta_1 = \text{diag}(0, 0, \ldots, 0)$, which corresponds to the subsystem when all sensors signals are off, we have that $F_{E,1} = A$. As a consequence, to guarantee the conditions of Lemma 3.1 it would be necessary for A to be Hurwitz stable, which cannot be always guaranteed in many applications. For this reason, we need to find a more general stability condition for switched systems in the presence of unstable subsystems and random switching.

The authors in [26], have introduced globally asymptotic stability conditions (GAS) for switched systems with stable and unstable subsystems, and where the switching signal has specific random properties. In fact, it only requires that the probability that the unstable subsystems are active to be small.

We are interested in the following definition of stability introduced in [26].

Definition 3.1. *The system (12) is said to be globally asymptotically stable almost surely (GAS a.s.) if the following two properties are simultaneously verified:*

$$Pr\left(\forall \epsilon > 0 \; \exists \beta > 0, \; \textit{such that } \|x_0\| < \beta \Longrightarrow \sup_{t \geq 0} \|x(t)\| < \epsilon)\right) = 1.$$

$$Pr\left(\forall r, \epsilon' > 0 \; \exists T \geq 0 \; \textit{such that } \|x_0\| < r \Longrightarrow \sup_{t \geq T} \|x(t)\| < \epsilon'\right) = 1$$

Definition 3.1 indicates that the solutions $t \mapsto x(t)$ converge to an equilibrium with probability 1 in finite time and from any bounded initial condition x_0.

The conditions for stability under random switching introduced in [26] employ a family of Lyapunov functions, one for each subsystem $F_{E,i}$ for $i \in \Sigma$, that possesses the following properties.

Assumption A1: There exist a family of continuously differentiable real-valued functions $V_i(x) \in \mathbb{R}$ for all $i \in \Sigma$, functions $\alpha_1, \alpha_2 \in \mathcal{K}_\infty$, numbers $\mu \geq 1, \lambda_i \in \mathbb{R}$ such that

(A1.1): $\alpha_1(\|x\|) \leq V_i(x) \leq \alpha_2(\|x\|), \ \ \forall x \in \mathbb{R}^n, \forall i \in \Sigma.$
(A1.2): $\dot{V}_i(x) \leq -\lambda_i V_i(x), \ \ \forall x \in \mathbb{R}^n, \forall i \in \Sigma,$
(A1.3): $V_i(x) \leq \mu V_j(x), \ \ \forall x \in \mathbb{R}^n, \forall i, j \in \Sigma.$

We also impose some assumptions on the switching signal.
Assumption A2: The switching signal $t \mapsto \sigma(t)$ satisfies the following properties:

- The sequence $(S_k)_{k\in\mathbb{N}}$, $S_{k+1} = t_{k+1} - t_k$ of holding times is a sequence of i.i.d. uniform random variables with parameter $T_{max} > 0$ and $T_{min} = 0$.
- The probability that the i^{th} subsystem is active is $\Pr(\sigma(t_k) = i) = \widetilde{p}_i$.
- S_k and σ_i are mutually independent.

The following result shows that Assumption A1 is satisfied for linear time-invariant systems.

Lemma 3.2. *For the linear time-invariant system $\dot{x} = H_i x$, with H_i not necessarily stable or unstable, the real-valued function $x \mapsto V_i(x) := x^\top Q_i x$, where $Q_i = Q_i^\top > 0$, satisfies Assumption A1 with*

$$\lambda_i \in \left\{ \lambda \in \mathbb{R} \ : \ H_i + \frac{\lambda}{2} I \text{ is Hurwitz} \right\}. \tag{13}$$

Proof: The first condition, (**A1.1**), can be easily verified since $V(x) = x^\top Q_i x$ is a convex quadratic function, such that there always exist class-$\mathcal{K}_\infty$ functions α_1 and α_2 defined as $\alpha_1(\|x\|) = \bar{\alpha}_1\|x\|^2$ and $\alpha_2(\|x\|) = \bar{\alpha}_2\|x\|^2$ for positive scalars $\bar{\alpha}_1 < \bar{\alpha}_2$ such that $\bar{\alpha}_1\|x\| \leq V_i(x) \leq \bar{\alpha}_2\|x\|$. For condition (**A1.2**), we have that $\dot{V}(x) = x^\top(H_i^\top Q_i + Q_i H_i)x$ for each x. Therefore, $\dot{V}(x) + \lambda_i V(x) < 0$ is equivalent to

$$x^\top \left[(H_i + \frac{1}{2}\lambda_i I)^\top Q_i + Q_i (H_i + \frac{1}{2}\lambda_i I) \right] x < 0. \tag{14}$$

Suppose that H_i is Hurwitz and let κ_i denote the eigenvalue with largest real part (i.e., the eigenvalue closest to imaginary axis). Recall that, by definition, κ_i satisfies $H_i v = \kappa_i v$, for v the corresponding eigenvector. For a scalar λ_i, the eigenvalues of $H_i + \frac{1}{2}\lambda_i I$ are such that $(H_i + \frac{1}{2}\lambda_i I)v = (\kappa_i + \frac{1}{2}\lambda_i)v$. Therefore, if λ_i satisfies $0 < \lambda_i < -2\text{Re}(\kappa_i)$, then $H_i + \frac{1}{2}\lambda_i I$ is Hurwitz and (14) holds. Similarly, if H_i is not stable with largest eigenvalue κ_i (further from zero), any $\lambda_i < -2\text{Re}(\kappa_i) < 0$ makes the term $H_i + \frac{1}{2}\lambda_i I$ Hurwitz and (14) is satisfied. The third assumption always holds for quadratic Lyapunov functions [26]. ■

The following theorem follows [26, Theorem 3.4], and introduces sufficient conditions for GAS a.s. of the switched system in (12) according to Definition 3.1.

Theorem 3.1. *Suppose that Assumptions **A1** and **A2** hold where $t \mapsto \sigma(t)$ has parameter T_{max} and probabilities $\widetilde{p}_i = \Omega(i)$ for all $i \in \Sigma$ according to (9). If*

$$\mu \sum_{i \in \Sigma} \widetilde{p}_i \left(\frac{1 - e^{-\lambda_i T_{max}}}{\lambda_i T_{max}} \right) < 1, \tag{15}$$

then the switched system is GAS a.s..

Theorem 3.1 is an adaptation of [26, Theorem 3.4] for linear systems, where $\sigma_i(t_k)$ follows the probability distribution Ω defined in (9). The proof is summarized in Appendix A

Remark 3.1. *If $\lambda_i < 0$, which is related to the unstable matrices, and it has large magnitude, the term $\left(\frac{1-e^{-\lambda_i T_{max}}}{\lambda_i T_{max}} \right)$ may be greater than one, such that the probability associated to that term has to be small enough; Therefore, in order to guarantee that (15) holds, $\widetilde{p}_i$ has to be chosen such that the unstable subsystems are selected with low probability.*

4. Detecting Stealthy Sensor Attacks

One of the main advantages of MTD is that it makes it harder for an adversary to tailor stealthy attacks due to the uncertainty added by the MTD mechanism. In particular, with our proposed sensor MTD, the adversary fails to predict how his attack affects the IDS, such that the attacks that are stealthy under normal conditions are visible with the MTD strategy.

We focus on a very powerful type of stealthy attack that has been introduced in [22, 23, 27]. Then, we show how, by appropriately selecting the probabilities p_i, it is possible to make these attacks visible, even when the adversary has access to the control inputs, all sensor readings, knows A, B, L, C, K, and knows the thresholds τ of the detection mechanism.

While we could try to define a similar non-stochastic defense by changing C deterministically, this would give the adversary more chances of finding the deterministic changes and adapt its attack accordingly. The uncertainty presented to the adversary is one of the advantages of MTD.

4.1. Construction of Stealthy Attacks

Suppose the attacker has access to all sensor readings and computes its own estimation of the system states $\hat{x}_a(t)$ in order to forge powerful cyber-attacks. The attacker's estimator is described by

$$\dot{\hat{x}}_a(t) = A\hat{x}_a(t) + Bu(t) + L(Cx(t) - C\hat{x}_a(t) + \delta^a(t)). \tag{16}$$

Let $s(t) = \hat{x}(t) - \hat{x}_a(t)$ denote the error between the system estimation used by the controller and the attacker estimation. We introduce the following lemma.

Lemma 4.1. *Suppose there is no MTD mechanism, i.e., $\Theta_{\sigma(t)} = I$, and L is such that $A - LC$ is Hurwitz. Then, the error $s(t)$ converges in the limit to $\lim_{t\to\infty} s(t) = 0$ and the attacker is able to compute an estimation that converges to the one used by the IDS.*

Proof: Notice that $\dot{s}(t) = \dot{\hat{x}}(t) - \dot{\hat{x}}_a(t)$. Combining (10) and (16) we get

$$\begin{aligned}\dot{s}(t) = F_{E,\sigma(t)}s(t) + L(\Theta_{\sigma(t)} - I)Ce(t)\\ - L(\Theta_{\sigma(t)} - I)Cs(t) + L(\Theta_{\sigma(t)} - I)\delta^a(t).\end{aligned} \tag{17}$$

Since $\Theta_{\sigma(t)} = I$, we have that $\dot{s}(t) = (A-LC)s(t)$, which is stable independently of $\delta^a(t)$ and the trajectories will always converge to 0. ■

In the remainder of this section we will assume that the system is in a steady state before the attack, such that $x(0) = 0, s(0) = 0$. These assumption will facilitate the derivation of the MTD design methodology but they are not necessary for the correct operation of the proposed MTD approach.

In the following lemma, we will introduce a type of stealthy attack that uses $\hat{x}_a(t)$ to bypass the IDS algorithm. This attack does not depend on the zero-dynamics, which makes it suitable for more general applications.

Lemma 4.2. *Suppose that the detection strategy corresponds to the bad-data detection introduced in (7) with detection thresholds $\tau = [\tau_1, \ldots, \tau_q]^\top$. If there is no MTD mechanism and the adversary launches an attack of the form*

$$\delta^a(t) = -y(t) + C\hat{x}_a(t) + \tau \tag{18}$$

then the attack remains stealthy.

Proof: Replacing (18) in (6), we obtain

$$\begin{aligned}r(t) &= C(x(t) - \hat{x}(t)) - C(x(t) - \hat{x}_a(t)) + \tau\\ &= -Cs(t) + \tau\end{aligned} \tag{19}$$

Without MTD, $s(t) = 0$ and the residuals are then $r(t) = \tau$. As a consequence, $h(t) = |\tau|$ and the alarm is never triggered. ■

Remark 4.1. *This type of attack is very powerful when the matrix A is not stable. If we apply the attack in* (18) *to* (12)*, the dynamics of the estimation error become $\dot{e}(t) = Ae(t) + L\Theta_{\sigma(t)}Cs(t) + L_{\sigma(t)}\tau$. If we define the extended state $w = [x^\top, e^\top, s^\top]^\top$, it is easy to see from $\dot{w} = \mathrm{Q}w + J\tau$ that part of the eigenvalues of* Q *correspond to the eigenvalues of A. If A is not stable, the attack causes the entire system to become unstable without being detected.*

4.2. Revealing Stealthy Attacks

We assume that the adversary does not know the MTD mechanism, such that he launches the stealthy attack in (18). The following theorem introduces the conditions to reveal the stealthy attack.

Theorem 4.1. *Suppose that the conditions in Theorem 3.1 are satisfied and an adversary launches the stealthy attack described in (18) for the bad-data detection strategy. Let* $E[\Theta_{\sigma(t)}] = \boldsymbol{P} = diag(p_1 \dots, p_q)$ *such that* $\bar{F}_E = A - L\boldsymbol{P}C$ *is Hurwitz. The stealthy attack is revealed if any of the following conditions holds for* at least one $j \in \mathcal{S}$,

$$\begin{aligned} &C_j \bar{F}_E^{-1} L(\boldsymbol{P} - I)\tau > 0, \\ &C_j \bar{F}_E^{-1} L(\boldsymbol{P} - I)\tau < -2\tau_j. \end{aligned} \tag{20}$$

Proof: Replacing the attack in (18) with the dynamics of the error in (17), we obtain

$$\dot{s}(t) = (A - L\Theta_{\sigma(t)}C)s(t) + L(\Theta_{\sigma(t)} - I)\tau. \tag{21}$$

Since τ is finite and constant, and since when $\delta^a(t) = 0$, (12) is GAS a.s. according to Theorem 3.1, then the term $L(\Theta_{\sigma(t)} - I)\tau$ will cause an accumulation of the error between the real effect of the attack and the effect estimated by the attacker. To facilitate the analysis, and since $s(t)$ is independent of $\sigma(t)$, we define $E[s(t)] = \bar{s}(t)$. Then, $\dot{\bar{s}}(t) = \bar{F}_E \bar{s}(t) + L(\boldsymbol{P} - I)\tau$. Therefore, the following limit exists

$$\lim_{t\to\infty} \bar{s}(t) = -\bar{F}_E^{-1} L(\boldsymbol{P} - I)\tau.$$

Applying the expectation operator $E[\cdot]$ to the residuals in (19) lead to

$$\lim_{t\to\infty} \bar{r}(t) = \left(C\bar{F}_E^{-1} L(\boldsymbol{P} - I) + I\right)\tau.$$

Given that the expected detection statistic is $E[h(t)] = |\bar{r}(t)|$, we have that

$$\lim_{t\to\infty} \bar{h}(t) = |C\bar{F}_E^{-1} L(\boldsymbol{P} - I)\tau + \tau|, \tag{22}$$

such that the attack is revealed when at least one $h_j(t) > \tau_j$, which is ensured if any of the conditions in (20) hold. ■.

Notice that the conditions for revealing the attack depend on $\boldsymbol{P}$. As a consequence, in the next section, we show how to impose constraints on $\boldsymbol{P}$ during the design process to guarantee that this type of strong stealthy attack is always revealed.

Remark 4.2. *With our proposed MTD, the attacker is not able to estimate the system states subject to his own attack. Therefore,* any attack that depends on the attacker's estimation can be potentially revealed. *Furthermore, as it was shown in the Motivation example in Section 2.6, other types of stealthy attacks can be also revealed.*

5. MTD Design

So far, Theorem 3.1 provides conditions for almost sure asymptotic stability of the system subject to the proposed MTD strategy. In Theorem 4.1 we derived conditions for revealing a powerful type of stealthy attacks. In this section we

introduce a methodology to design the probability matrix $\boldsymbol{P}$ that: 1) ensures that the stealthy attacks introduced in Section 4 are revealed, 2) reduces the impact of most attack trajectories using an input-to-state stability (ISS) criteria, and 3) guarantees stability and desired performance constraints.

Recall that $\boldsymbol{P} = E[\Theta(t)] = \text{diag}(p_1, p_2, \ldots, p_q)$, and let $\bar{z}(t) = E[z(t)] = [\bar{x}^\top(t), \bar{e}^\top(t)]^\top$. Applying the expectation operator $E[\cdot]$ to (12), we obtain

$$\dot{\bar{z}}(t) = \bar{F}\bar{z}(t) + \bar{G}\bar{\delta}^a(t), \tag{23}$$

where $E[\delta^a(t)] = \bar{\delta}^a(t)$, and

$$\bar{F} = \begin{bmatrix} A+BK & -BK \\ 0 & A-L\boldsymbol{P}C \end{bmatrix}, \ \bar{G} = \begin{bmatrix} 0 \\ -L\boldsymbol{P} \end{bmatrix}.$$

Impact of the Attack: We can define the impact of the attack in terms of how much any attack trajectory can deviate the system trajectories in expectation. To this end, we will utilize input-to-state-state stability (ISS) criteria. Let us consider the expected dynamic system in (23). When $\bar{F}$ is Hurwitz, the state trajectories can be bounded as follows [28]:

$$|\bar{z}(t)| \leq \beta|\bar{z}(0)| + \gamma|\bar{\delta}_a|_\infty, \ \ \forall t \geq 0, \tag{24}$$

where $\beta = \kappa e^{-\alpha(\bar{F})}$, $\gamma = \frac{\|\bar{G}\|}{\alpha(\bar{F})}$, and $\alpha(\bar{F}) := \min\{|Re(\lambda)| : \lambda \in eig(\bar{F})\}$. Notice that the effect of an attack depends specifically on γ. Therefore, in order to reduce the impact of any bounded attack trajectory, it is possible to find $\boldsymbol{P}$ that minimizes γ.

Remark 5.1. *If $\bar{\delta}^a(t)$ is constant, we have that*

$$\lim_{t\to\infty} \bar{z}(t) = -\bar{F}^{-1}\bar{G}\bar{\delta}^a, \tag{25}$$

leading to $\lim_{t\to\infty} \bar{x}(t) = (A+BK)^{-1}BK(A-L\boldsymbol{P}C)^{-1}L\boldsymbol{P}\bar{\delta}^a$. Therefore, we can quantify the impact of the attack as follows:

$$\mathcal{I}(\boldsymbol{P}, \bar{\delta}^a) = \|M\bar{\delta}^a\|,$$

for $M = (A+BK)^{-1}BK(A-L\boldsymbol{P}C)^{-1}L\boldsymbol{P}$. The impact $\mathcal{I}(\boldsymbol{P}, \bar{\delta}^a)$ is then reduced by finding $\boldsymbol{P}$ that minimizes $\gamma = \|M\|$.

Performance under MTD: It is necessary to design an MTD that, not only reveals stealthy attacks and decreases the impact of any attack trajectory, but also guarantees some performance conditions in the attack-free case (e.g., convergence speed).

In order to quantify the degradation caused by the MTD mechanism in the system, we use as a performance index the slowest eigenvalue of $\bar{F}_E = A - L\boldsymbol{P}C$, which is related to the convergence speed of the observer. Therefore, our goal is to design an MTD strategy that guarantees

$$\lambda_{max}(A - L\boldsymbol{P}C) \leq \widetilde{\lambda},$$

where $\widetilde{\lambda} < 0$ and $\lambda_{max}(A - LC) < \widetilde{\lambda} < 0$.

Finally, since we want to find $\boldsymbol{P}$ to reveal stealthy attacks according to Theorem 4.1, let $\Psi_j^+ = C_j \bar{F}_E^{-1} L(\boldsymbol{P} - I)\tau$ and $\Psi_j^- = C_j \bar{F}_E^{-1} L(\boldsymbol{P} - I)\tau + 2\tau_j$. Thus, we can define

$$\Psi = \sum_{j \in \mathcal{S}} \max\{\Psi_j^+, 0\} - \min\{\Psi_j^-, 0\},$$

such that *at least one* $h_j > \tau_j$ when $\Psi > 0$.

The following optimization problem (Problem OMTD) allows us to find, for given $\widetilde{\lambda}$ and τ, the MTD probabilities associated to each sensor that guarantee GAS a.s., ensure that the type of stealthy attacks introduced in 18 are revealed, and minimizes the impact of the attack.

Problem OMTD

$$\begin{aligned}
&\min_{\boldsymbol{P}} \gamma \\
&s.t. \\
&0 < p_j \leq 1, \ \forall j \in \mathcal{S} \\
&\Psi > 0, \\
&(9), (15) \\
&\lambda_{max}(\bar{F}_E) \leq \widetilde{\lambda}.
\end{aligned} \tag{26}$$

Note that this is a nonlinear optimization problem that, at times, can be solved using interior-point or active-set algorithms.

5.1. Simplified Optimization Problem

The optimization problem in (26) possesses non-convex constraints that can increase the difficulty to find a solution, specially due to the probability distribution in (9) and the term $\bar{F}_E^{-1}$. For this reason, we will reformulate Problem OMTD by taking advantage of some properties of the probability distribution that maps $\boldsymbol{P}$ into $\widetilde{p}$, and we will simplify some of the constraints using upper and lower bounds. In this case, we will focus our attention on finding $\widetilde{p} \in \mathbb{R}^s$ instead of matrix $\boldsymbol{P}$. The only limitation arises when the number of sensors q increases because the size of $\widetilde{p}$ increases exponentially according to $s = 2^q$.
Recall that we have the set of non-identical binary matrices $\{\Theta_1, \ldots, \Theta_s\}$. Now, let $\boldsymbol{\Theta} = [\Theta_1 \ \Theta_2 \ \ldots \ \Theta_s]$ be a $q \times (sq)$ matrix formed by all matrices Θ_i. Also, recall that $\widetilde{p}_i$ denotes the probability that Θ_i is true. It is easy to see that $p_i = \sum_{j=1}^{p} \widetilde{p}_j [\Theta_i]_j$, which is the converse of (9). As a consequence, we have that, for $\widetilde{p} = [\widetilde{p}_1, \ldots, \widetilde{p}_s]^\top$,

$$\boldsymbol{P} = \text{diag}(\boldsymbol{\Theta}(\widetilde{p} \otimes \mathbf{1_q})) \tag{27}$$

Clearly, the relationship between $\widetilde{p}$ and P is linear. Also, the condition in (15) is linear with respect to $\widetilde{p}_i$, such that, for $v_i = \mu\left(\frac{1-e^{-\lambda_i T_{max}}}{\lambda_i T_{max}}\right)$ and for

$v = [v_1, \ldots, v_s]^\top$, we have that (15) is equivalent to $v^\top \widetilde{p} < 1$.
Other complexity that can be simplified is the inverse terms in γ and in Ψ. First, recall that under the stealthy attack in 18, detection statistic is $h_i(t) = |r_i(t)| = |-C_i s(t) + \tau_i|$. Therefore, if we make the term $-C_i s(t)$ sufficiently large or small, then it is possible to make the stealthy attack detectable. To this end, let $\|C_i s(t)\|^2 = s(t)^\top C_i^\top C_i s(t)$. Also, recall that $\bar{s} = -\bar{F}_E^{-1} L(\boldsymbol{P} - I)\tau$. Therefore, using the Rayleigh-Ritz Inequality we obtain

$$\begin{aligned} \|C_i \bar{s}\| &= \tau^\top (\boldsymbol{P} - I) L^\top {F_E^{-1}}^\top C_i^\top C_i F_E^{-1} L(\boldsymbol{P} - I)\tau \\ &\geq \lambda_{min}(C_i^\top C_i)\lambda_{min}({F_E^{-1}}^\top F_E^{-1})\tau^\top (\boldsymbol{P} - I) L^\top L(\boldsymbol{P} - I)\tau \\ &= \lambda_{min}(C_i^\top C_i)\frac{1}{\|\bar{F}_E\|^2}\|L(\boldsymbol{P} - I)\tau\|^2 \end{aligned}$$

Therefore, finding $\boldsymbol{P}$ that maximizes $\|L(\boldsymbol{P}-I)\tau\|^2$ and minimizes $\|\bar{F}_E\|$ can lead to revealing stealthy attacks. Similarly, the objective function associated with γ can be rewriten as a minimization of $\|\bar{G}\| - \alpha(\bar{F})$. The new objective function for the design optimization problem is now $(\|F_E\|^2 - \|L(\boldsymbol{P}-I)\tau\|^2) - \varsigma(\|\bar{G}\| - \alpha(\bar{F}))$ where $\varsigma > 0$ defines the priority given to attack detection or to attack minimization. The simplified observer-based MTD design problem (SOMTD) is then:

Problem SOMTD

$$\begin{aligned} \min_{\widetilde{p}} \ & (\|F_E\|^2 - \|L(\boldsymbol{P} - I)\tau\|^2) - \varsigma(\|\bar{G}\| - \alpha(\bar{F})) \\ & s.t. \\ & \widetilde{p}^\top \mathbf{1}_{\boldsymbol{n}} = 1 \\ & \boldsymbol{P} = \mathrm{diag}(\boldsymbol{\Theta}(\widetilde{p} \otimes \mathbf{1}_{\boldsymbol{q}})) \\ & 0 < \widetilde{p}_i \leq 1, \ \forall i \in \Sigma \\ & v^\top \widetilde{p} < 1 \\ & \lambda_{max}(\bar{F}_E) \leq \widetilde{\lambda}. \end{aligned}$$

The existence of a solution is conditioned according to the following Lemma.

Lemma 5.1. *Consider Problem SOMTD for given constants $\varsigma, \widetilde{\lambda}$, and detection thresholds τ. If $\widetilde{\lambda} > \lambda_{max}(A - LC)$, then there exists at least one local minimizer that satisfies the constraints.*

Proof: The proof is easily verified given that there always exists a combination of $\widetilde{p}_i$ that satisfy constraints (1)-(4). Given that any $\boldsymbol{P}$ different to the identity would lead to slower eigenvalues, γ can never be smaller than the non MTD case. ■

6. Case Studies

In order to verify the viability of the proposed algorithms, we consider two case studies: i) a generic LTI system and ii) a vehicular platooning problem

with 10 vehicles. The MTD design for each case will be performed using the optimization problems introduced in Section 4.

6.1. LTI Dynamic System

We consider an continuous-time MIMO LTI system described by the following matrices

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

The feedback control gain K computed by solving the LQR problem with unitary costs and the steady state Kalman filter gain are given by

$$K = \begin{bmatrix} -6.2 & -1.23 & -0.77 \\ -4 & -0.88 & -0.35 \end{bmatrix}, \quad L = \begin{bmatrix} 2.0726 & 0.3431 \\ 0.2040 & 0.5216 \\ 0.1312 & 0.1362 \end{bmatrix}.$$

The MTD design parameters are$\gamma = -1$, $T_{max} = 0.1$ and the detection thresholds $\tau = [0.02, 0.05]^\top$, and they are used for testing the two proposed MTD design problems.

For the first case, the solution of Problem OMTD is found using an interior-point algorithm and corresponds to $\boldsymbol{P}^* = \text{diag}([0.98, 0.62])$. Figure 2 illustrates the trajectories in normal operation of the sensor signal y_2 the estimated state $\hat{x}_2$, and the system state x_2 with and without the optimal MTD. Notice that the effects of switching are significantly smoothed by the proposed MTD observer introduced in (10). In particular, small transients can be observed in the state estimation dynamics but they are less evident in the system states.

Now, for comparison, an arbitrary selection of $\boldsymbol{P} = \text{diag}([0.7, 0.5])$ is also considered. Figure 3 illustrates the Montecarlo simulation of the trajectories of the states and the norm $\|x(t)\|$ for the attack $\delta^a(t) = [-0.2, 2]^\top$ after 20 s. Clearly, the optimal MTD approach is able to decrease the impact of this attack when compared to the case without MTD. Given that $\gamma = -1$, which is close to $\lambda_{max}(A - LC)$ (the maximum eigenvalue without MTD), the performance degradation is minimal, and the convergence with and without MTD are very similar. The arbitrary MTD causes a significant degradation in the performance and increases the impact of the attack when compared with the case without MTD, illustrating the importance of the proposed MTD design.

Now, suppose that an adversary launches a stealthy attack as described in (18) for the anomaly detection threshold τ. Figure 4 shows the detection metrics h_1, h_2 introduced in (7). Without MTD, the attack remains completely stealthy . However, thanks to the random MTD mechanism, the attack is now revealed. Notice that by using arbitrary switching probabilities, the detection capabilities increase due to lower probabilities p_1, p_2 when compared to the optimal MTD case; however, as it was illustrated in Fig. 3, the performance is significantly degraded. The proposed optimal design maintains an adequate balance between detection and performance.

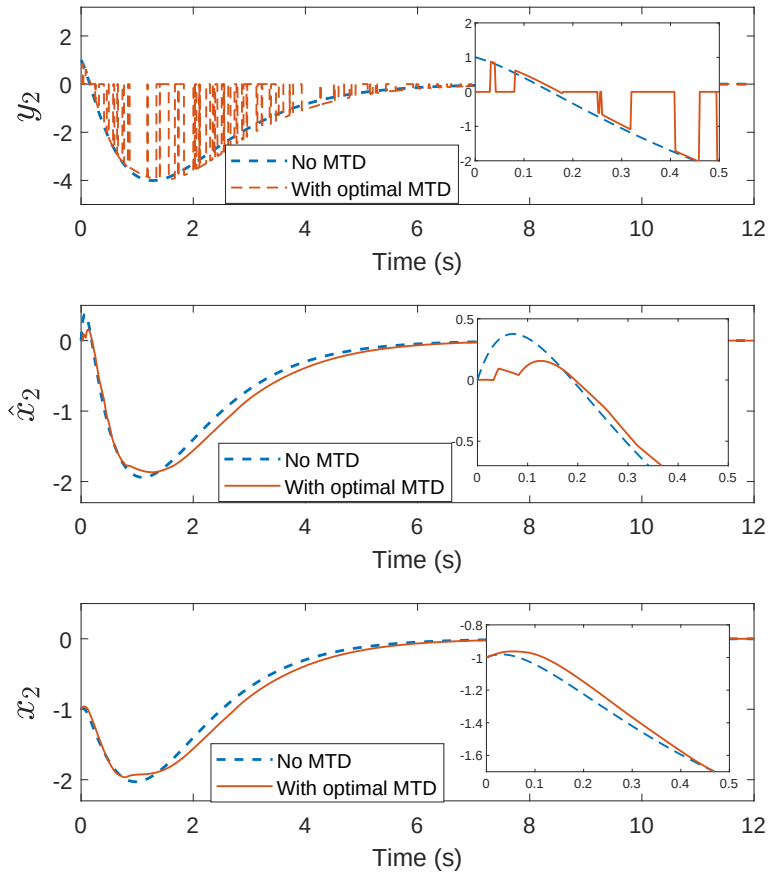

Figure 2: y_2, $\hat{x}_2$, and x_2 in normal operation with and without MTD. The proposed MTD induces transient that are smoothed through the MTD observer. Even though small transients can be observed in the state estimation, they are hardly evident in the system states.

For the second case, the MTD design using Problem SOBMD depends on the selection of α and γ. Figure 5 depicts the probabilities of each sensor for different values of α and γ. When $\alpha = 0$, the optimization problem will focus only on increasing the detection of the stealthy attack, without minimizing the attack impact. On the other hand, large α will give priority to the minimization of the attack impact and the solution may not detect the stealthy attacks. Notice also that larger γ leads to a less conservative MTD due to an increase in the permitted performance degradation. By contrasting both optimization problems, we can see that the solution of Problem OMTD is equivalent to the solution of Problem SOMTD when α is large, such that for this particular case, focusing on attack impact minimization inherently lead to stealthy attack detection.

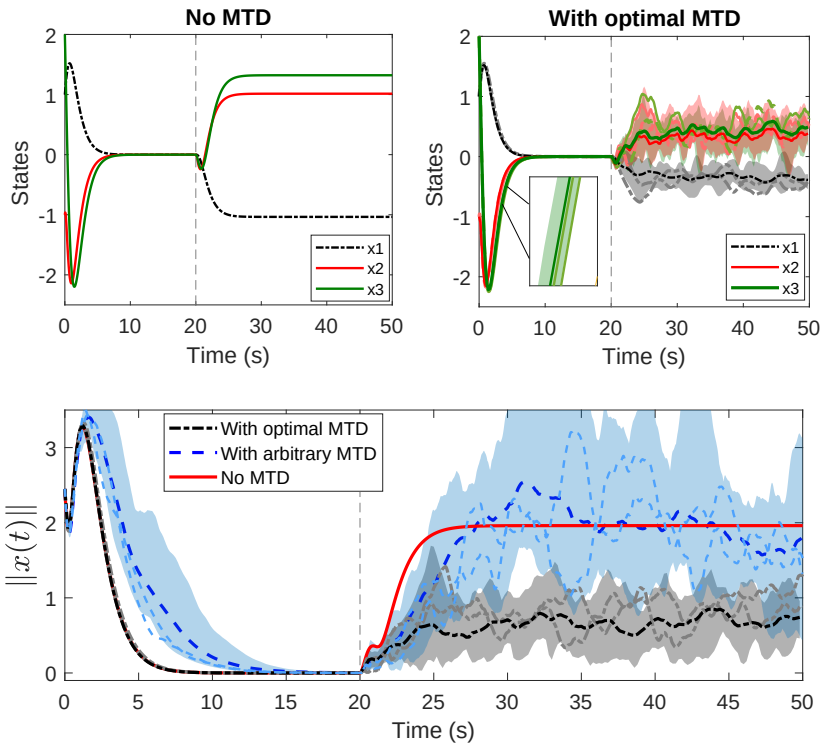

Figure 3: System states without MTD (top left), with the MTD strategy (top right), and the norm of the states $\|x(t)\|$ (bottom). The shaded area indicate the maximum/minimum deviation of the Montecarlo simulation at each time instant. Notice that our approach decreases the deviation caused by the attack, and since $\gamma = -1$ and $p_1 = 0.98$, the performance degradation is small. An arbitrary selection of switching probabilities $\boldsymbol{P} = \text{diag}([0.7, 0.5])$ can significantly degrade the system performance.

6.2. Vehicular Platooning

In this second test case, we consider a system of n_v cooperating autonomous vehicles that form a vehicular platoon [29], as illustrated in Figure 6. We consider that each vehicle uses on-board sensors (e.g., lidar) to maintain a given distance with its immediate neighbors [30]. In addition, the platoon possesses a cooperative controller modeled by an additive acceleration term, which is computed by a centralized cloud that collects sensor measurements transmitted through wireless communications. The dynamics of the positions x_i and velocities v_i of the vehicles are described with the following differential equations [31],

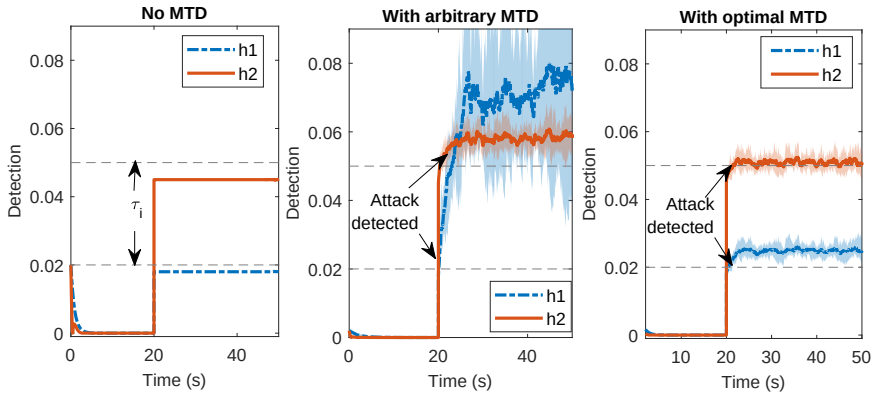

Figure 4: Bad-data detection with thresholds $\tau = [0.02, 0.05]^\top$ in the presence of a stealthy attack where h_1, h_2 denote the detection score introduced in (7). The attack is never detected without MTD (left). The addition of uncertainty makes possible to reveal the attack (center, right). The arbitrary MTD increases the capability to detect the stealthy attack but at the cost of performance degradation as depicted in Fig. 3

$$
\begin{cases}
\dot{\mathrm{x}}_1 = v_1 \\
\vdots \\
\dot{\mathrm{x}}_{n_v} = v_{n_v} \\
\dot{v}_1 = \quad k_p(\mathrm{x}_2 - \mathrm{x}_1 - d^*) + k_d(v_2 - v_1) + \beta v_1 + u_1 \\
\dot{v}_2 = -k_p(\mathrm{x}_2 - \mathrm{x}_1 - d^*) - k_d(v_2 - v_1) \\
\qquad +k_p(\mathrm{x}_3 - \mathrm{x}_2 - d^*) + k_d(v_3 - v_2) + \beta v_2 + u_2 \\
\vdots \\
\dot{v}_{k-1} = -k_p(\mathrm{x}_{k-1} - \mathrm{x}_{k-2} - d^*) - k_d(v_{k-1} - v_{k-2}) \\
\qquad +k_p(\mathrm{x}_{n_v} - \mathrm{x}_{n_v-1} - d^*) + k_d(v_k - v_{k-1}) + \beta v_{k-1} + u_{k-1} \\
\dot{v}_k = -k_p(\mathrm{x}_k - \mathrm{x}_{k-1} - d^*) - k_d(v_k - v_{k-1}) + \beta v_k + u_k
\end{cases}
\tag{28}
$$

where $k_p = 2$ and $k_d = 1.5$ are the proportional and derivative gains of an on-board Proportional-Derivative (PD) controller, which regulates the distance between neighboring vehicles to be the desired distance $d^* = 2$ m; $\beta = -0.1$ characterizes the loss of velocity as a result of friction; and u_i with $i \in \{1, \ldots, n\}$ are feedforward inputs (acceleration) added to each vehicle. In cooperative cruise control settings, such feedforward inputs are used to optimize the performance of the platoon by each vehicle sharing its intended maneuvers, thus requiring the PD control to only compensate for errors. The platoon is most concisely described by the *relative* distances between each pair of adjacent vehicles, defined as $d_{i,i+1} = x_{i+1} - x_i$, for $i = 1 \ldots, n_v - 1$. We can introduce new relative distance error variables $e_{i,i+1} = d_{i,i+1} - d^*$, such that in the equilibrium, $e_{i,i+1} = 0$ implies $d_{i,i+1} = d^*$. Considering $x = [d_{1,2}, \ldots d_{n_v-1,n_v}, v_1, \ldots v_n]^\top$, we can rewrite (28) in terms of $2n_v - 1$ state variables (i.e., $n_v - 1$ relative distance errors and n_v velocities) such that $\dot{x}(t) = Ax(t) + Bu(t)$, with input $u = [u_1, \ldots, u_{n_v}]^T$, $B = [0, I]^T$. Let us consider the case when $n_v = 10$, and

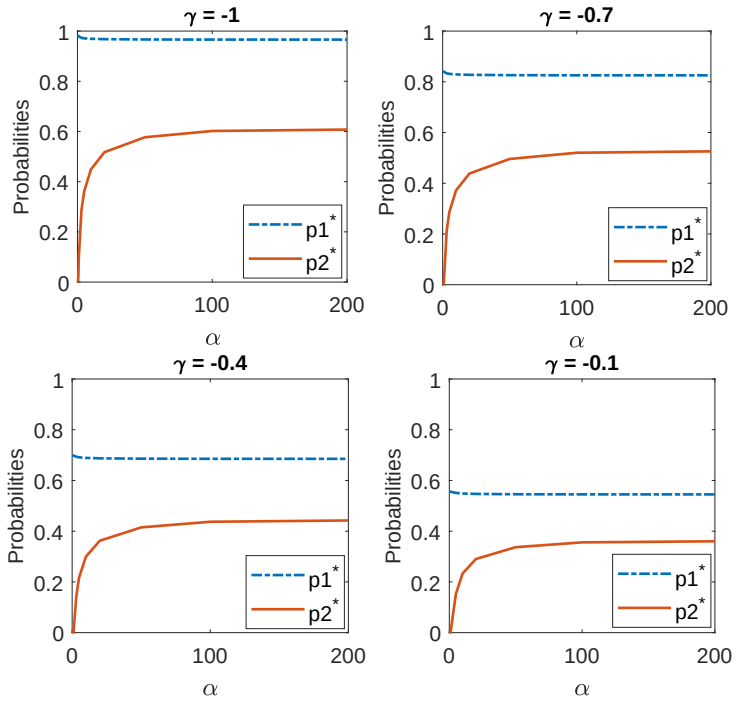

Figure 5: MTD design by solving Problem SOBMD probabilities for each sensor for different values of α and γ

suppose that each vehicle only transmits the velocities to the cooperative controller, such that the number of sensors is $q = 10$ and $s = 1024$. Given that s is large, and since A is stable, we can use Problem SBOMD with $\gamma = -0.2$, $\tau_i = \tau = 2$. The solution is then $\boldsymbol{P}^* = \text{diag}([1, 1, 1, 1, 1, 1, 0.9594, 0.65, 0.3, 0.1])$.

The goal of an attacker is to cause any of pair of vehicles to crash. We assume the attacker can only manipulate the sensor measurements sent by the vehicle n_v, which is the one in front, and inject a bias of 1 m/s. Figure 7(left) depicts the impact of the attack and how it causes a crash in multiple vehicles. However, with the proposed MTD approach, the same attack is no longer successful, as shown in the Montecarlo simulation depicted in Fig. 7(right) for 100 simulations. Figure 8 illustrates how the same attack in different vehicles have different impacts. Clearly, the proposed design algorithms can successfully associate lower probabilities to the sensors that are most sensitive to attacks.

7. Conclusions

We have proposed and analyzed the security of an MTD strategy for improving the detectability of attacks, while at the same time minimizing the power that an adversary has when compromising a sensor signal. We showed that

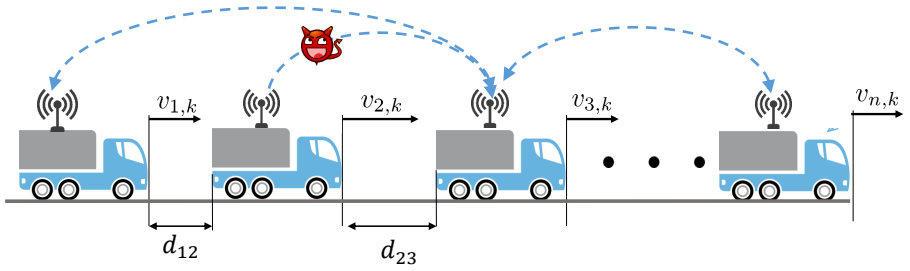

Figure 6: Illustration of a vehicular platooning with cooperative autonomous cruise control.

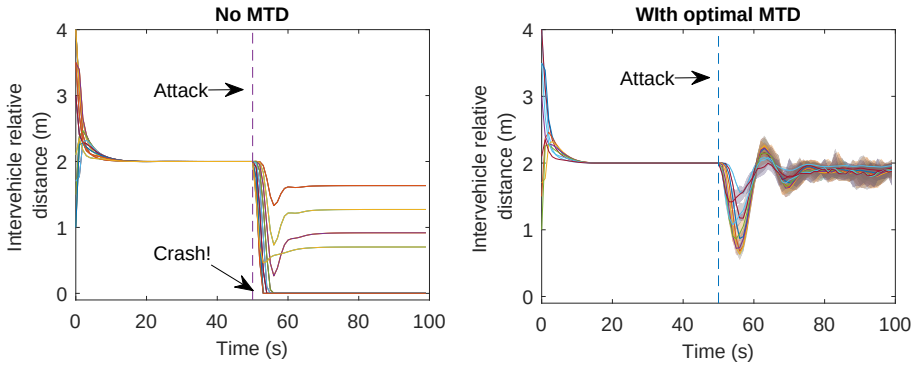

Figure 7: Example of a bias attack in vehicle 10. Without MTD the attack is able to cause a crash between several vehicles (left). A Monte-Carlo simulation shows that with MTD the impact of the attack is considerably reduced.

our strategy is effective against very powerful stealthy attacks even when the adversary knows the system dynamics, the detection strategy, and has access to all sensors and control inputs. We derived conditions for global stability of the system and defined an optimization problem that allows us to find the probability at which each sensor transmits in such a way the state deviation caused by the attack is minimized while guaranteeing the detection of stealthy attacks. A simplified optimization problem was proposed in order to facilitate the computation of the optimal solution. We evaluated the proposed design approaches in two case studies and illustrated the benefits of our proposed MTD. In practice the MTD strategy can be activated when we notice indicators of attacks, or if we notice that the system is deviating from the desired space without explanation; if the MTD is activated then, it will be able to mitigate the attack while at the same time revealing a previously undetected attack.

8. Acknowledgements

Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-20-1-0253. The views and conclusions contained

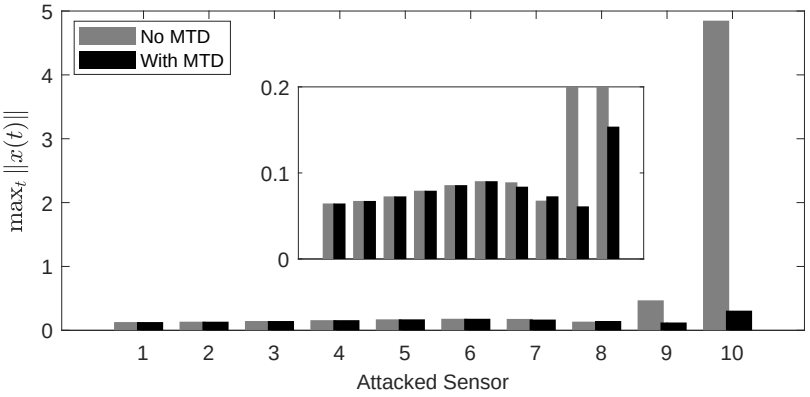

Figure 8: Maximum deviation of $\|x(t)\|$ for a single bias attack of 1 m/s in different sensors. Notice that without MTD, attacking sensor 10 causes the largest deviation. The optimal MTD design assigns the lowest probability to sensor 10.

in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein

References

[1] Y. Liu, M. K. Reiter, P. Ning, False data injection attacks against state estimation in electric power grids, in: ACM Conference on Computer and Communications Security, Chicago, IL, USA, 2009, pp. 21–32.

[2] A. Teixeira, S. Amin, H. Sandberg, K. H. Johansson, S. Sastry, Cyber security analysis of state estimators in electric power systems, in: IEEE Conf. on Decision and Control, Atlanta, GA, USA, 2010, pp. 5991–5998.

[3] S. Bhattacharya, T. Başar, Differential game-theoretic approach to a spatial jamming problem, in: Advances in Dynamic Games, Springer, 2013, pp. 245–268.

[4] S. Maharjan, Q. Zhu, Y. Zhang, S. Gjessing, T. Başar, Dependable demand response management in the smart grid: A stackelberg game approach., IEEE Trans. Smart Grid 4 (1) (2013) 120–132.

[5] H. Fawzi, P. Tabuada, S. Diggavi, Secure estimation and control for cyber-physical systems under adversarial attacks, IEEE Transactions on Automatic Control 59 (6) (2014) 1454–1467.

[6] C.-Z. Bai, F. Pasqualetti, V. Gupta, Security in stochastic control systems: Fundamental limitations and performance bounds, in: American Control Conference, Chicago, Il, 2015, pp. 195–200.

[7] F. Pasqualetti, F. Dörfler, F. Bullo, Attack detection and identification in cyber-physical systems, IEEE Transactions on Automatic Control 58 (11) (2013) 2715–2729.

[8] S. Sundaram, C. Hadjicostis, Distributed function calculation via linear iterative strategies in the presence of malicious agents, IEEE Transactions on Automatic Control 56 (7) (2011) 1495–1508.

[9] S. Phillips, A. Duz, F. Pasqualetti, R. G. Sanfelice, Hybrid attack monitor design to detect recurrent attacks in a class of cyber-physical systems, in: Proceedings of the 2017 IEEE Conference on Decision and Control, NULL, 2017, pp. 1368–1373.

[10] A. Duz, S. Phillips, A. Fagiolini, R. G. Sanfelice, F. Pasqualetti, Stealthy attacks in cloud-connected (linear-impulsive) systems, in: Proceedings of the American Control Conference, 2018, pp. 146–152.

[11] S. Jajodia, A. K. Ghosh, V. Swarup, C. Wang, X. S. Wang, Moving target defense: creating asymmetric uncertainty for cyber threats, Vol. 54, Springer Science & Business Media, 2011.

[12] E. O. of the President, Trustworthy cyberspace: Strategic plan for the federal cyber security research and development program, Tech. rep., National Science and Technology Council (2011).

[13] K. R. Davis, K. L. Morrow, R. Bobba, E. Heine, Power flow cyber attacks and perturbation-based defense, in: Proceedings of the IEEE Third International Conference on Smart Grid Communications (SmartGridComm), 2012, IEEE, 2012, pp. 342–347.

[14] M. A. Rahman, E. Al-Shaer, 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, ACM, 2014, pp. 59–68.

[15] J. Tian, R. Tan, X. Guan, T. Liu, Hidden moving target defense in smart grids, in: Proceedings of the 2nd Workshop on Cyber-Physical Security and Resilience in Smart Grids, ACM, 2017, pp. 21–26.

[16] P. Griffioen, S. Weerakkody, B. Sinopoli, A moving target defense for securing cyber-physical systems, IEEE Transactions on Automatic Control (2020) 1–1doi:10.1109/TAC.2020.3005686.

[17] S. Weerakkody, B. Sinopoli, Detecting integrity attacks on control systems using a moving target approach, in: Proceedings of the IEEE 54th Annual Conference on Decision and Control (CDC),, IEEE, 2015, pp. 5820–5826.

[18] A. Kanellopoulos, K. G. Vamvoudakis, Entropy-based proactive and reactive cyber-physical security, in: Proactive and Dynamic Network Defense, Springer, 2019, pp. 59–83.

[19] J. Tian, R. Tan, X. Guan, Z. Xu, T. Liu, Moving target defense approach to detecting stuxnet-like attacks, IEEE Transactions on Smart Grid 11 (1) (2020) 291–300.

[20] D. Umsonst, H. Sandberg, On the confidentiality of controller states under sensor attacks, Automatica 123 (2021) 109329.

[21] J. Giraldo, M. El Hariri, M. Parvania, Moving target defense for cyber-physical systems using iot-enabled data replication, IEEE Internet of Things Journal (2022).

[22] J. Giraldo, A. Cardenas, M. Kantarcioglu, Security and privacy trade-offs in cps by leveraging inherent differential privacy, in: 2017 IEEE Conference on Control Technology and Applications (CCTA), 2017, pp. 1313–1318.

[23] C. Murguia, J. Ruths, Cusum and chi-squared attack detection of compromised sensors, in: 2016 IEEE Conference on Control Applications (CCA), 2016, pp. 474–480.

[24] J. Giraldo, A. Cardenas, R. G. Sanfelice, A moving target defense to reveal cyber- attacks in cps and minimize their impact, in: Proceedings of the American Control Conference, 2019, pp. 391–396.

[25] H. Lin, P. J. Antsaklis, Stability and stabilizability of switched linear systems: a survey of recent results, IEEE Transactions on Automatic control 54 (2) (2009) 308–322.

[26] D. Chatterjee, D. Liberzon, Stabilizing randomly switched systems, SIAM Journal on Control and Optimization 49 (5) (2011) 2008–2031.

[27] D. I. Urbina, J. A. Giraldo, A. A. Cardenas, N. O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, H. Sandberg, Limiting the impact of stealthy attacks on industrial control systems, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, CCS '16, ACM, New York, NY, USA, 2016, pp. 1092–1105.

[28] H. K. Khalil, Nonlinear systems; 3rd ed., Prentice-Hall, Upper Saddle River, NJ, 2002.

[29] Y. Zheng, S. E. Li, J. Wang, L. Y. Wang, K. Li, Influence of information flow topology on closed-loop stability of vehicle platoon with rigid formation, in: 17th International IEEE Conference on Intelligent Transportation Systems (ITSC), IEEE, 2014, pp. 2094–2100.

[30] K. C. Dey, A. Rayamajhi, M. Chowdhury, P. Bhavsar, J. Martin, Vehicle-to-vehicle (v2v) and vehicle-to-infrastructure (v2i) communication in a heterogeneous wireless network–performance evaluation, Transportation Research Part C: Emerging Technologies 68 (2016) 168–184.

[31] S. Dadras, R. M. Gerdes, R. Sharma, Vehicular platooning in an adversarial environment, in: Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security, ACM, 2015, pp. 167–178.

Appendix A. Proof of Theorem 3.1

For $t \in [\tau_i, \tau_{i+1}]$, and $S_{i+1} = \tau_{i+1} - \tau_i$ we have from **A1.2** and **A1.3** that

$$\begin{aligned} V_{\sigma(\tau_{i+1})}(x(\tau_{i+1})) &\leq V_{\sigma(\tau_{i+1})}(x(\tau_i))e^{-\lambda_{\sigma(\tau_i)}S_{i+1}} \\ &\leq \mu V_{\sigma(\tau_i)}(x(\tau_i))e^{-\lambda_{\sigma(\tau_i)}S_{i+1}} \end{aligned}$$

Iterating the above inequality and employing **A1.1** we obtain

$$V_{\sigma(\tau_i)}(x(\tau_i)) \leq \alpha_2(\|x(\tau_0)\|)\prod_{j=0}^{i-1}\mu e^{-\lambda_{\sigma(\tau_j)}S_{j+1}}$$

Employing the independence hypothesis in **A2** we have that

$$E[V_{\sigma(\tau_i)}(x(\tau_i))] \leq \alpha_2(\|x(\tau_0)\|)\prod_{j=0}^{i-1}\mu E\left[e^{-\lambda_{\sigma(\tau_j)}S_{j+1}}\right]. \tag{A.1}$$

Given that S_{j+1} is drawn from an uniform distribution, then

$$\begin{aligned} E\left[e^{-\lambda_{\sigma(\tau_j)}S_{j+1}}\right] &= E\left[E_{s_{j+1}}\left[e^{-\lambda_{\sigma(\tau_j)}S_{j+1}}\right]\right] \\ &= E\left[\int_0^{T_{max}} \frac{1}{T_{max}} e^{-\lambda_{\sigma(\tau_i)}s} ds\right] \\ &= E\left[\frac{1-e^{-\lambda_{\sigma(\tau_i)}T_{max}}}{\lambda_{\sigma(\tau_i)}T_{max}}\right] \\ &= \sum_{l\in\Sigma}\widetilde{p}_l\left(\frac{1-e^{-\lambda_l T_{max}}}{\lambda_l T_{max}}\right) \end{aligned}$$

Substituting in (A.1), leads to

$$E[V_{\sigma(\tau_i)}(x(\tau_i))] \leq \alpha_2(\|x_0\|)\left(\mu\sum_{j\in\Sigma}\widetilde{p}_j\left(\frac{1-e^{-\lambda_j T_{max}}}{\lambda_j T_{max}}\right)\right)^i.$$

Clearly, if the condition in (15) holds, then $\lim_{i\to\infty} E[V_{\sigma(\tau_i)}(x(\tau_i))] = 0$. Now, based on these results, it is possible to follow the steps in [26] to show how by satisfying (15), conditions in Definition (3.1) are satisfied. ■