Hybrid Design of Multiplicative Watermarking for Defense Against Malicious Parameter Identification

Zhang, J.; Gallo, Alexander J.; Ferrari, Riccardo M.G.

DOI
10.1109/CDC49753.2023.10383837

Publication date
2023

Document Version
Final published version

Published in
Proceedings of the 62nd IEEE Conference on Decision and Control (CDC 2023)

2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Hybrid Design of Multiplicative Watermarking for Defense Against Malicious Parameter Identification

Jiaxuan Zhang*, Alexander J. Gallo* and Riccardo M. G. Ferrari*

Abstract— **Multiplicative watermarking (MWM) is an active diagnosis technique for the detection of highly sophisticated attacks, but is vulnerable to malicious agents that use eavesdropped data to identify and then remove or replicate the watermark. In this work, we propose a scheme to *protect* the parameters of MWM, by proposing a design strategy based on piecewise affine (PWA) hybrid dynamical systems, called hybrid multiplicative watermarking (HMWM). Due to the design decision to make certain states of the HMWM systems unobservable, we show that parameter reconstruction by an eavesdropper is infeasible, from both a computational and a system-theoretic perspective, while not altering the system's closed-loop performance.**

Attack Detection, Cyber-Physical Security, Resilient Control Systems

I. Introduction

Modern Industrial Control Systems (ICS) often employ Information Technology (IT) hardware and software, in order to be more performant and reach greater interoperability. This evolution has exposed critical industrial infrastructures to cyber attacks [1], with them compromising information between controller and plant, and thus possibly leading to system-level disruption. *The design of secure CPSs is therefore imperative*, and *secure control* has emerged as an active research area [2].

A promising direction corresponds to *active attack detection methods* [3]–[8], which do not rely only on plant dynamics knowledge, but actively modify inputs or outputs to enhance attack detectability. Often, these methods rely on matched mechanisms being present on both the plant and the controller side of communication networks to generate and then validate or remove the additional signals. Although active detection methods have been shown to improve detection capabilities against malicious agents injecting false data, they do so under the assumption that attackers do not adapt their behaviour in response to the defence strategies. Indeed, if the attacker successfully identifies the additional security measures put in place for defence, the injected data can be suitably adapted to evade detection. Different methods have been proposed as countermeasures to this, e.g., in [5], [6], [9], [10], the parameters of the active diagnosis scheme are switched or generated over time, thus changing the parameters that must be identified by an attacker to remain stealthy. In [6], [9], new parameters are generated using pseudo-random number generators, one at the plant and one at the controller side, which must be synchronized to guarantee proper performance. A switching mechanism is proposed in [5], relying on an event-triggered strategy to define when to update the parameters of the multiplicative watermarking. In [11] a method based on the elliptic curve cryptography is proposed to further improve security for switching multiplicative watermarking (MWM).

Several techniques exist in literature to counteract the presence of malicious eavesdropping attacks in ICS communication networks. Solutions based on differential privacy [12], [13] rely on injecting additional noise to the data to ensure sensitive information cannot be recovered through eavesdropping. This however comes at the detriment of overall performance. Other techniques exist to encrypt controllers [14], though at the cost of additional computation time, and therefore impacting the stability margin.

In this paper, we propose a novel active diagnosis method based on MWM, such that the watermarking filters are explicitly designed to resist identification. Specifically, we propose a *hybrid multiplicative watermarking* (HMWM), where the watermark generator and remover are defined as piecewise affine (PWA) hybrid dynamical systems, with unobservable states. Compared to existing MWM schemes [5], we prove that our proposed design ensures that parameter reconstruction by an eavesdropping attack remains infeasible.

The main contributions of this paper are:

- a hybrid multiplicative watermarking method for the security of CPSs, based on PWA hybrid dynamical systems with unobservable states;
- an algorithm for their design, demonstrating that the obtained HMWM scheme does not alter the stability or performance of the closed-loop CPS;
- the method, by exploiting the HMWM systems' switching dynamics, is shown to resist identification by attackers;
- an example of a switching function under which each mode is active with uniform probability.

The rest of this paper is organized as follows. In Section II, we formulate the problem, by defining the system structure, the MWM filters, and the attacker capabilities. In Section III, we propose the PWA HMWM system design for the filters, presenting the algorithm to be followed for parameter design. In Section IV, we analyze the HMWM method's performance in resisting the identification from eavesdropping attackers. Finally, in Section V we demonstrate our scheme's effectiveness via numerical simulations. To satisfy space

*Delft University of Technology, Delft, The Netherlands. {j.zhang-42@student., a.j.gallo@, r.ferrari@}tudelft.nl

This paper has been partially supported by the AIMWIND project, which is funded by the Research Council of Norway under grant no. 312486.

2023 62nd IEEE Conference on Decision and Control (CDC) | 979-8-3503-0124-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CDC49753.2023.10383837

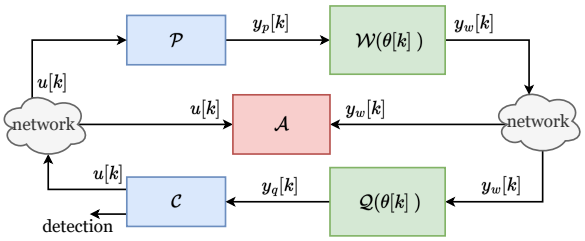


Fig. 1: Complete CPS: in blue the plant and controller, in red the attacker, in green the multiplicative watermarking pair.

constraints, we omit our proofs, which may however be found in [15].

Notation: $\mathbb{Z}_+$ denotes the set of nonnegative integers. I_n represents the n-dimensional identity matrix, while $0_{n\times m} \in \mathbb{R}^{n\times m}$ is a matrix of zeros; whenever clear from context, the subscripts n and $n \times m$ are omitted. Given a matrix $X \in \mathbb{R}^{n\times n}$, $\sigma(X)$ denotes its spectrum, and $\rho(X)$ its spectral radius. A matrix $X \in \mathbb{R}^{n\times n}$ is said to be orthogonal, or orthonormal if it is invertible and $X^{-1} = X^\top$. For any two matrices X_1 and X_2, let $X = \mathrm{diag}(X_1, X_2)$ denote the block-diagonal matrix defined by X_1 and X_2. The notation $X \succ (\succeq) 0$ is used to state that a symmetric matrix $X \in \mathbb{R}^n$ is positive (semi)definite; similarly, for negative (semi)definite matrices, $x \prec (\preceq) 0$. For a time-varying signal $x[k] \in \mathbb{R}^n$, $k \in \mathbb{Z}_+$, $x[k_1 : k_2]$ is the sequence of instances $x[k], k \in \{k_1, k_1 + 1, \ldots, k_2\} \subseteq \mathbb{Z}_+$. A polyhedron $\mathscr{X} \subset \mathbb{R}^{n\times n}$ is a convex set, defined as $\mathscr{X} = \{x \in \mathbb{R}^n : Hx \leq k\}$, where $H \in \mathbb{R}^{m\times n}$ and $k \in \mathbb{R}^m$. For any two sets $\mathscr{A}$ and $\mathscr{B}$, $\mathscr{A}\times\mathscr{B}$ denotes their Cartesian product. We use $x \sim \mathcal{N}(\mu, \Sigma)$ to define a normally distributed random variable x with mean μ and variance Σ.

II. System Description and Problem Formulation

We consider a cyber-physical system composed of a physical plant $\mathcal{P}$ and a controller $\mathcal{C}$. The information between the controller and plant is exchanged over a communication network: this exposes the CPS to an eavesdropping attacker $\mathcal{A}$, capable of reading input and output signals. To counteract this, we suppose the CPS is equipped with a switching multiplicative watermarking pair $(\mathcal{W}, \mathcal{Q})$. The considered CPS structure is shown in Figure 1.

A. System Model

The plant is modelled as an LTI system with dynamics

$$\mathcal{P}: \begin{cases} x_p[k+1] &= A_p x_p[k] + B_p u[k] + w_p[k]; \\ y_p[k] &= C_p x_p[k] + v_p[k] \end{cases} \tag{1}$$

where $x_p \in \mathbb{R}^n, y_p \in \mathbb{R}^p$ are the plant's state and measurement output, and $u[k] \in \mathbb{R}^m$ is the control input. The signals $w_p \in \mathbb{R}^n$ and $v_p \in \mathbb{R}^p$ represent process and measurement noise, assumed to be i.i.d. Gaussian $w_p[k] \sim \mathcal{N}(0, \Sigma_w)$, $v_p[k] \sim \mathcal{N}(0, \Sigma_v)$. Furthermore, the initial condition is $x_p[0] \sim \mathcal{N}(x_{p,0}, \Sigma_0)$. We assume that all matrices are of appropriate dimensions, and that (A_p, B_p) and (C_p, A_p) are respectively controllable and observable pairs. Controller $\mathcal{C}$ regulates the system, and performs anomaly detection[1]. Specifically, it is implemented as the following LTI system:

$$\mathcal{C}: \begin{cases} \hat{x}_p[k+1] = A_p\hat{x}_p[k] + B_p u[k] + L(y_p[k] - C_p\hat{x}_p[k]) \\ u[k] = -K(\hat{x}_p[k] - x_{p,ref}) + u_{ref} \\ r[k] = y_p[k] - C_p\hat{x}_p[k] \end{cases} \tag{2}$$

where $\hat{x}_p \in \mathbb{R}^n$ is the estimated state, with $\hat{x}_p(0) = x_{p,0}$, $x_{p,ref} \in \mathbb{R}^n$, $u_{ref} \in \mathbb{R}^m$ are the reference state and control input, which are assumed to be piecewise constant. Matrices L and K are designed such that $A_p - LC_p$ and $A_p - B_pK$ are stable. $r \in \mathbb{R}^p$ is used as a residual for attack diagnosis: the specific definition of the diagnosis tool is omitted from this paper, as out of its scope, but interested readers can turn to [5] for further details.

B. Multiplicative Watermarking

Proposed in [5], [16], switching multiplicative watermarking is an active technique for attack detection, whereby a watermarking generator ($\mathcal{W}$) *filters* y_p before its transmission over the communication network to the controller. Once received, a suitably defined watermarking remover ($\mathcal{Q}$) then processes the information, returning a signal used by the controller. Specifically, $\mathcal{W}$ and $\mathcal{Q}$ are time-varying systems, designed to have the following dynamics:

$$\begin{aligned} \mathcal{W}: &\begin{cases} x_w[k+1] = A_w(\theta_w[k])x_w[k] + B_w(\theta_w[k])y_p[k] \\ y_w[k] = C_w(\theta_w[k])x_w[k] + D_w(\theta_w[k])y_p[k] \end{cases} \\ \mathcal{Q}: &\begin{cases} x_q[k+1] = A_q(\theta_q[k])\, x_q[k] + B_q(\theta_q[k])\, y_w[k] \\ y_q[k] = C_q(\theta_q[k])\, x_q[k] + D_q(\theta_q[k])\, y_w[k] \end{cases} \\ \mathcal{F}: &\ \theta_w[k] = f_w(\mathcal{I}_w[k]); \quad \theta_q[k] = f_q(\mathcal{I}_q[k]) \end{aligned} \tag{3}$$

where $x_w, x_q \in \mathbb{R}^{n_w}$ are the generator and remover states, $y_w, y_q \in \mathbb{R}^p$ their outputs, and $\theta_w[k], \theta_q[k] \in \mathbb{R}^{n_\theta}$ their parameters at time k, with $n_\theta = (n_w + p)^2$; $f_w : \mathscr{I}_w \to \mathbb{R}^{n_\theta}$ and $f_q : \mathscr{I}_q \to \mathbb{R}^{n_\theta}$ are switching functions.

Definition 1 (Watermarking pair)**.** *Two systems $(\mathcal{W}, \mathcal{Q})$, with dynamics* (3)*, are called a watermarking pair if:*

a. $\mathcal{W}$ and $\mathcal{Q}$ are stable and invertible;
b. if $\theta_w[k] = \theta_q[k]$, $y_q[k] = y_p[k]$, i.e., $\mathcal{Q} = \mathcal{W}^{-1}$. ◁

To meet Definition 1.a., the matrices for $\mathcal{Q}$ are defined as:

$$\begin{aligned} &D_q(\theta) = D_w(\theta)^{-1}; A_q(\theta) = A_w(\theta) - B_w(\theta)D_q(\theta)C_w(\theta); \\ &B_q(\theta) = B_w(\theta)D_q(\theta); \quad C_q(\theta) = -D_q(\theta)C_w(\theta). \end{aligned} \tag{4}$$

where $\theta = \theta_w[k] = \theta_q[k]$. Any MWM design method must ensure that this condition is satisfied for all k; to do so the information available at the MWM generator and remover at time k, $\mathcal{I}_w$ and $\mathcal{I}$ is used, where:

$$\begin{aligned} \mathscr{I}_w[k] &\triangleq \{y_w[0:k], y_p[0:k], x_w[0:k], \theta_w[0:k]\}, \\ \mathscr{I}_q[k] &\triangleq \{y_w[0:k], y_q[0:k], x_q[0:k], \theta_q[0:k], u[0:k]\}. \end{aligned} \tag{5}$$

[1]Note that, although not the focus of this paper, we have included an anomaly detector, as multiplicative watermarking is predominantly a method for active attack diagnosis.

C. Attacker Capabilities

We define the following threat model for the eavesdropper attacker $\mathcal{A}$ depicted in Fig. 1:

System knowledge: The attacker knows the parameters of the plant and controller models $\{A_p, B_p, C_p, L, K\}$; it is also aware that a MWM scheme is present on the CPS.

Disclousure resources: The attacker has direct access to signals y_w and u transmitted over the communication network. The set of information available to the attacker at time k can be therefore defined as:

$$\mathscr{I}_a[k] \triangleq \{A_p, B_p, C_p, L, K, u[0:k], y_w[0:k]\}\,. \quad (6)$$

Note that $\theta_w[0], \theta_q[0] \notin \mathscr{I}_a$.

Attack objective: The malicious agent attempts to reconstruct the MWM parameters $\theta_w[k]$ and $\theta_q[k]$ for all $k \geq K_{id}^a, k \in \mathbb{Z}_+$. Without loss of generality, $K_{id}^a = 0$.

D. Problem Formulation

The switching rules represented by f_w and f_q affect the difficulty for a malicious agent to identify the parameters of the multiplicative watermarking parameters. The switching rules we are to design should meet the following requirements:

- **R1 Fast Switching**: The mode must switch rapidly;
- **R2 Randomness**: The switching sequence must not be known in advance;
- **R3 Synchronization**: $\mathcal{W}$ and $\mathcal{Q}$ must have synchronized modes, i.e., the mode must be chosen based on common information of $\mathscr{I}_w[k]$ and $\mathscr{I}_q[k]$.

Remark 1. *We set requirement **R1** to avoid any design strategy that includes a minimum dwell time, as it has been shown to be beneficial for parameter identification, as is pointed out in Section IV. This does not lead to undesirable behavior, such as Zeno behavior, as we consider discrete-time systems.* ◁

Given the scenario presented in the previous subsections, the problem this paper addresses can be formalized as follows:

Problem 1. *Given a cyber-physical system* (1)-(2)*, equipped with a multiplicative watermarking scheme* (3)*, design the time-varying parameters θ_w, θ_q such that:*

a. $(\mathcal{W}, \mathcal{Q})$ is a watermarking pair, as per Definition 1;
b. the CPS maintains closed-loop stability under switching;
*c. an attacker with the information set $\mathscr{I}_a$ and capabilities defined in Section II-C cannot exactly reconstruct $\theta_w[k], \theta_q[k]$, for all $k \geq K_{id}^a$, i.e. the time and data complexity to exactly identify the parameters can be arbitrarily large. In this paper, it relates to meeting the requirements **R1**-**R3**.* ◁

III. Design of Hybrid Multiplicative Watermarking

A. HMWM Structure

We propose a design strategy that defines the dynamics of $\mathcal{W}$ and $\mathcal{Q}$ as piecewise affine (PWA) linear switched systems. More precisely, the dynamics of $\mathcal{W}$ are[2]:

$$\mathcal{W}: \begin{cases} x_w[k+1] = \sum_{i=0}^{N} \beta_{w,i} \left(A_{w,i} x_w[k] + B_{w,i} y_p[k]\right) \\ y_w[k] = \sum_{i=0}^{N} \beta_{w,i} \left(C_{w,i} x_w[k] + D_{w,i} y_p[k]\right) \end{cases} \quad (7)$$

$$\mathcal{F}: \theta_w[k] = \sum_{i=0}^{N} \beta_{w,i} \theta_i\,,\ \beta_{w,i} = \begin{cases} 1, \text{ if } x_{w,u}[k] \in \mathscr{P}_i \\ 0, \text{ otherwise} \end{cases}$$

where subscript $i \in \mathcal{N} = \{1, \ldots, N\}$ indicates one of N *modes* of operation, and the boolean variables $\beta_{w,i}[k] \in \{0,1\}, \forall i \in \mathcal{N}$ are used to determine which mode is active at any given time. The mode of the system is determined by evaluating in which set $\mathscr{P}_i$ the state $x_{w,u} \in \mathbb{R}^{n,u}$ belongs. $x_{w,u}$, defined explicitly in the following, is a portion of x_w which is unobservable from y_w, by design. These are polyhedric sets which cover $\mathbb{R}^{n_u}$. To avoid any ambiguity caused by the non-zero intersection of neighboring subsets, we introduce the following huristic, guaranteeing that $\sum_{i \in \mathcal{N}} \beta_{w,i} = 1$: if $x_{w,u} \in \mathscr{P}_i \cap \mathscr{P}_j$, $\beta_{w,i} = 1$ iff $i < j$. An example of these polyhedrons is shown in Fig. 2a.

Let us now proceed with our proposed design method, summarized in Alg. 1. To guarantee that $\mathcal{W}$ is stable under arbitrary switching, admits a stable inverse, and is unobservable, we define matrices in (7) as follows:

$$A_{w,i} = \begin{bmatrix} A_{w,i}^- & 0 \\ 0 & A_{w,u} \end{bmatrix}, \quad B_{w,i} = \begin{bmatrix} B_{w,i}^- \\ B_{w,u} \end{bmatrix}, \quad C_{w,i} = \begin{bmatrix} C_{w,i}^- & 0 \end{bmatrix}, \quad D_{w,i} = D_{w,i}^-\,. \quad (8)$$

This definition leads to the definition of the unobservable state, as x_w can be partitioned as $x_w = \left[x_{w,o}^\top, x_{w,u}^\top\right]^\top$, where $x_{w,u}$ is unobservable by design. We define $A_{w,u} = \mathrm{diag}(a_{w,1}, \ldots, a_{w,n_u})$ and $B_{w,u} \in \mathbb{R}^{n_u \times p}$ to be common to all modes, and bound $|a_{w,j}| \leq \sqrt{0.5}, \forall j \in \{1, \ldots, n_u\}$ to guarantee stability. Matrices $A_{w,i}^-$ are defined as $A_{w,i}^- \triangleq \bar{T}^\top \bar{A}_{w,i}^- \bar{T}$, where $\bar{A}_{w,i}^-$ are randomly defined, stable, diagonal matrices for all $i \in \mathcal{N}$, and $\bar{T}$ is an orthogonal matrix common to all modes. Matrices $B_{w,i}^-$ are defined such that the pair $(A_{w,i}^-, B_{w,i}^-)$ is stabilizable. Matrices $C_{w,i}^-$ and $D_{w,i}^-$ are then defined as follows: firstly, a matrix K_i stabilizing $A_{w,i}^- - B_{w,i}^- K_i$ is found satisfying

$$\begin{bmatrix} X & A_{w,i}^- X + B_{w,i}^- Z_i \\ (A_{w,i}^- X + B_{w,i}^- Z_i)^\top & X \end{bmatrix} \succ 0 \quad X \succ 0; \quad K_i = -Z_i X^{-1} \quad \forall i \in \mathcal{N}; \quad (9)$$

$D_{w,i}^-$ is defined to be random, square and invertible, and finally $C_{w,i}^-$ satisfies $C_{w,i}^- = D_{w,i} K_i$. This procedure guarantees that $(\mathcal{W}, \mathcal{Q})$ is a watermarking pair.

[2]The dynamics of $\mathcal{Q}$ are analogous to (7), substituting subscript w with q, changing $y_p[k]$ to $y_w[k]$, and defining the system matrices following (4).

Algorithm 1 Generate GUAS $\mathcal{W}$ and $\mathcal{Q}$

Input: $n_w \geq 1, N \geq 1$
Output: $\theta_i, i \in \mathcal{N}$

1: Randomly generate diagonal matrices $\bar{A}^-_{w,i}$, $i \in \mathcal{N}$, such that $\rho(\bar{A}^-_{w,i}) < 1$, and an orthonormal matrix $\bar{T}$
2: Define $A^-_{w,i} = \bar{T}^\top \bar{A}^-_{w,i} \bar{T}$.
3: Randomly generate $B^-_{w,i}$ such that $(A^-_{w,i}, B^-_{w,i})$ are controllable;
4: Design K_i such that (9) is jointly satisfied for all $i \in \mathcal{N}$;
5: Randomly generate $D_{w,i}$ and define $C_{w,i} = D_{w,i} K_i$.
6: Randomly generate $a_w \in \mathbb{R}$ & $|a_w| \leq \sqrt{0.5}, b_w^\top \in \mathbb{R}^p$, set $A_{w,u} = a_w$ and define $A_{w,i}, B_{w,i}, C_{w,i}, D_{w,i}$ solving (8).
7: Define $A_{q,i}$, $B_{q,i}$, $C_{q,i}$, $D_{q,i}$, corresponding to $A_{w,i}$, $B_{w,i}$, $C_{w,i}$, $D_{w,i}$, solving (4);
8: **if** $n_w > 1$, **for** $t = 2 : n_w$, define:

$$A^-_{w,i} = A_{w,i}\,, \quad B^-_{w,i} = B_{w,i}\,,$$
$$C^-_{w,i} = C_{w,i}\,, \quad D^-_{w,i} = D_{w,i}\,;$$

9: Repeat Step 6
10: **endif endfor**

B. Requirement satisfaction

Throughout this subsection, we show that the proposed algorithm generates a stable watermarking pair under arbitrary switching, the states of which remain synchronized, and which do not alter the performance of the closed-loop CPS. Furthermore, we show that the requirements for the MWM design established in Prob. 1.a. and Prob. 1.b. in Sec. II-D are met by the precedure defined in Alg. 1.

Theorem 1. *Given a watermark generator and remover pair $(\mathcal{W}, \mathcal{Q})$ with dynamics as in* (4)*, if their system matrices are generated following Algorithm 1, with $n_u \geq 1$, the systems are GUAS and input-state stable (ISS) under arbitrary switching. Furthermore, $\mathcal{Q}(\theta_i) = \mathcal{W}(\theta_i)^{-1}$, $\forall i \in \mathcal{N}$.* □

Proposition 1. *Suppose CPS in* (1)-(2) *is equipped with the HMWM scheme* (7)*. If $x_w[0] = x_q[0]$, and $\mathcal{W}$ and $\mathcal{Q}$ share the same $\mathscr{P}_i, \forall i \in \mathcal{N}$, then $\theta_w[k] = \theta_q[k], \forall k \geq 0$.* □

Proposition 2. *The closed-loop of the CPS with watermarking pair $(\mathcal{W}, \mathcal{Q})$ designed following Alg. 1 is stable, and its performance remains unchanged, if $x_w[0] = x_q[0]$.* □

Having shown that stability holds, we now discuss that the requirements outlined in Section II-D are satisfied. Indeed, **R1** is met because, although exact quantification of the dwell time between switching events is challenging, the boundaries of each region $\mathscr{P}_i$ can be defined such that the probability of $x_{w,u}[k] \in \mathscr{P}_i$ is uniform across all $i \in \mathcal{N}$, given knowledge of the probability distributions of $w[k]$ and $v[k]$; **R2** is met, as the switching can be seen as being "truly random": the dynamics of $x_{w,u}$ depend on w and v, which are the result of physical processes, and are not generated by a pseudo-random number generator[3]. Therefore, it is not possible to define the trajectory of $x_{w,u}[k]$ *a priori*. Finally, Prop. 1 proves state and parameter synchronization, fulfilling **R3**.

Remark 2. *Let us note here that the switching law we present in this paper is different to the one presented in [5] in one fundamental aspect. Indeed, here the switching law at time k depends on information available in $\mathscr{I}_w[k-1]$ and $\mathscr{I}_q[k-1]$. Instead, in [5], the authors propose an event-triggered switching law, and the watermark remover must first decode $y_w[k]$, then evaluate whether there has been a parameter jump in the watermark generator, and if that is the case, update its own parameters and recompute $y_q[k]$.* ◁

C. Example Design of Switching Region

Let us now propose a possible definition of the partitions $\mathscr{P}_i, i \in \mathcal{N}$. Specifically, we propose a partitioning of $\mathbb{R}^{n_u}$ such that, when the system reaches steady state, the probability of $x_{w,u}[k] \in \mathscr{P}_i$, at any k, is uniform across $i \in \mathcal{N}$. Let us start by characterizing the statistical properties of $x_{w,u}[k] \sim \mathcal{N}(\mu_{x_{w,u}}[k], \Sigma_{x_{w,u}}[k])$. From dynamics in (7), we obtain:

$$\begin{aligned}\mu_{x_{w,u}}[k] &= A_{w,u}\mu_{x_{w,u}}[k-1] + B_{w,u}\mu_{y_p}[k-1]\\ \Sigma_{x_{w,u}}[k] &= A_{w,u}\Sigma_{x_{w,u}}[k-1]A_{w,u}^\top + B_{w,u}\Sigma_{y_p}[k-1]B_{w,u}^\top\end{aligned} \tag{10}$$

where $\mu_{y_p}[k]$ and $\Sigma_{y_p}[k]$ are the mean and variance of $y_p[k]$, which can be characterized by $\mu_{x_p}[k]$ and $\Sigma_{x_p}[k]$, and in turn be characterized by the estimation error $e[k] = x_p[k] - \hat{x}_p[k]$. Assume the controller uses a steady-state Kalman filter gain L [17], given that $A_p - B_p K$ is Schur stable, it is possible to define the steady state values of $\mu_{x_{w,u}}$ and $\Sigma_{x_{w,u}}$. The steady-state statistics of $x_{w,u}$ can then be used to partition $\mathbb{R}^{n_u}$ into N polyhedra, each having the same probability, using, e.g., the cumulative distribution function of the multiparametric Gaussian distribution.

Remark 3. *The procedure outlined in this section only considers using $x_{w,u}[k]$ for mode selection. This is, of course, only one possible solution, as mode selection can also depend on $x_w[k]$ as a whole, or $u[k]$.* ◁

Remark 4. *Note that the procedure proposed in this section to define $\mathscr{P}_i, i \in \mathcal{N}$ depends on the references $x_{p,ref}, u_{ref}$. As such, it is necessary to change $\mathscr{P}_i$ whenever $x_{p,ref}$ changes. We leave the development of a definition of the partitioning $\mathscr{P}_i$ that is time-invariant as future work.* ◁

IV. Identification Resistance

We now turn to evaluate whether our proposed method satisfies Prob. 1.c.. In this section, we are inspired by evaluation methods for cryptographic algorithms. Indeed, from the perspective of cryptography, $\mathcal{W}$ and $\mathcal{Q}$ can be seen as procedures to encode and decode transmitted data. $\theta_w[k]$ and $\theta_q[k]$ can be thus seen as secret keys, guaranteeing security.

[3]Note that at design stage random-number generators are necessary for the definition of the system parameters; this is done offline and does not clash with our statement here.

In assessing the security of cryptographic algorithms, the computational complexity required to *break* them is evaluated. This often takes the form of evaluating the complexity of solving inverse problems [18]. The techniques for evaluating the security of cryptographic algorithms inspire our analysis, relying on three metrics:

i. the *computational complexity* of identifying the system parameters;
ii. the amount of *memory* required to perform identification;
iii. an evaluation of the *theoretical difficulties* associated with identifying a PWA model with unobservable states.

Theorem 2. *Considering $\mathcal{W}$ and $\mathcal{Q}$, designed following Algorithm 1, the computational complexity of exactly identifying $\theta_w[k]$ and $\theta_q[k]$ from $\mathscr{I}_a[k]$ is $\mathcal{NP}$-hard.* □

Remark 5. *In [19, Ch.5], the basis of the proof of Thm. 2, analysis of the complexity of different bounded-error identification strategies for switched systems is conducted by restricting solutions to the set of rational numbers. We apply these results here without loss of generality, as in practice the solution we propose is likely to be applied to a digital control system with fixed point representation.* ◁

Although Theorem 2 gives a result for the computational complexity of *exact* identification of the system parameters, there are some methods to find some *approximate* solutions for input-output (IO) models of the system, such as the piecewise auto-regressive model with extra input (PWARX). One possible method is piecewise affine regression [19]. The following result pertains to the difficulty of identifying PWA systems with unobservable outputs.

Theorem 3. *An HMWM system $\mathcal{W}$ and $\mathcal{Q}$ designed following Algorithm 1, does not admit a PWARX model.* □

Remark 6. *Some literature present methods to identify state-space models directly, e.g., [19]–[22]. However, [21], [22] assume a minimum dwell time and [20] requires the system to be pathwise-observable, these requirements are not satisfied by the scheme presented in this paper.* ◁

Thm. 3 shows that there do not exist exact finite dimensional IO reprepresentations of $\mathcal{W}$ and $\mathcal{Q}$ resulting from Alg. 1, do to their construction as unobservable systems. It may still be possible to define a suitable finite-dimensional IO model to approximate the switching dynamics. Such an approximation requires a minimum number of data points, to ensure persistency of excitation (PE) [23]. Thus the adversary must have sufficient physical memory resources to store this data. In Tab. I we give the amount of data required (sample complexity) and dimension of the IO model, supposing n_m modes and a horizon n_h [23], [24]. As n_m and n_h grow, the sample complexity becomes untractable.

TABLE I: IO Identification Complexity

IO	IO dimension	Sample Complexity
$n_m^{n_h}$	$(p+m)n_h$	$\frac{((p+m)n_h-1)n_m^{n_h}+((p+m)n_h+1)n_m^{n_h}}{2}$

V. Numerical Example

A. Simulation setup

We use the linearized quadruple-tank water system in [8] as our test bench. The noise parameter, the linearized operating points, and the controller parameters we used are as follows:

$$x_{\text{ref}} = [5, 5, 2.044, 1.399]^\top, u_{\text{ref}} = [0.724, 1.165]^\top$$
$$\mu_w = [0,0,0,0]^\top, \mu_v = [0,0]^\top, \Sigma_w = 10^{-3} I_4, \Sigma_v = 10^{-1} I_2$$
$$K = \begin{bmatrix} -3.0993 & -4.0721 & 2.0528 & -2.8417 \\ -3.9353 & -3.3330 & -2.8461 & 1.9997 \end{bmatrix}$$

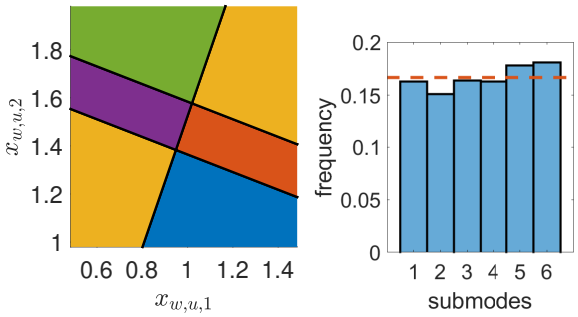
$\mathcal{W}$ and $\mathcal{Q}$ are designed with $n_w = 5$ and $n_u = 2$, and $N = 6$. To test the performance of our technique, we randomly generate 50 sets of parameters in this way, and for each set we run a simulation for 1000 steps. As an example, one set of parameters of the hybrid multiplicative watermarking generator's unobservable state is as follows:

$$A_{w,u} = \begin{bmatrix} 0.3908 & 0 \\ 0 & 0.6076 \end{bmatrix}, B_{w,u} = \begin{bmatrix} 0.1299 & 0.4694 \\ 0.5688 & 0.0119 \end{bmatrix}$$

Following the procedure in Sec. III-C, we partition $\mathbb{R}^{n_u}$ as in Fig. 2a, with steady-state mean and variance (10):

$$\mu_{x_{w,u}} = \begin{bmatrix} 0.9838 \\ 1.4800 \end{bmatrix}, \Sigma_{x_{w,u}} = \begin{bmatrix} 0.0287 & 0.0105 \\ 0.0105 & 0.0535 \end{bmatrix}$$

For these watermarking parameters, the IO model dimension and the minimum number of samples needed to meet the PE requirement for different horizon lengths are given in Table II. As expected, the number of IO models and samples needed becomes intractable as the horizon number grows.



(a) (b)

Fig. 2: (a) partitioning of $\mathbb{R}^{n_u}, n_u = 2$ into $N = 6$ polyhedrons, each associated to a parameter $\theta_i, i \in \mathcal{N}$; (b) relative frequency of each of the $N = 6$ modes over 1000 time steps.

B. Performance

The simulation results demonstrate that, indeed, the requirements in Sec. III-A are met. **R1**: throughout the simulation there are an average of 844 switching events, with a median and a maximum dwell time of 1 and 6, respectively; furthermore, in Fig. 2b we show the sample distribution of each mode, which as desired approaches the uniform distribution. **R2**: this requirement is guaranteed by design, as discussed in Sec. III, because $x_{w,u}$ depends on y_p, and

TABLE II: IO Identification Complexity

Horizon Number	IO modes	number of samples
5	7776	5.7451×10^8
10	6.0466×10^7	7.1295×10^{16}
15	4.7018×10^{11}	6.5217×10^{24}

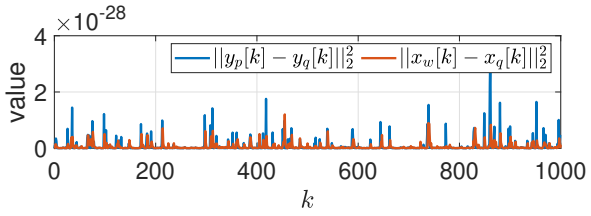


Fig. 3: Switching Synchronization: absolute difference between outputs of $\mathcal{P}$ and $\mathcal{Q}$, and states of $\mathcal{W}$ and $\mathcal{Q}$.

therefore on the exact realization of w and v. **R3**: as shown in Fig. 3, the error between y_p and y_q, as well as that between x_w and x_q, remains negligible. This error can be ascribed to numerical errors in MATLAB.

Let us now evaluate our proposed method's efficacy against identification by an eavesdropping attack. We suppose the attacker attempts to estimate the labels of the system modes by using input-output data, and implementing two methods available in literature [25], namely k-means and k-LinReg. In Tab. III we show that, for different indices (the random index (RI), the Fowlkes-Mallows index (FMI), and the Jaccard index (JI)), for different horizon lengths, the attacker's clustering method is ineffective.

VI. Conclusion

In this work, we propose a piecewise-affine hybrid design for multiplicative watermarking. We provide methods to design parameters which guarantee the multiplicative watermarking systems are stable and invertible, and present a way of partitioning the state space such that the resulting switching is fast and randomic, while maintaining synchronization. We demonstrate the hardness that our proposed methodology offers against eavesdropping attacks.

In the future, we will focus on the detection performance of the proposed method, as well as investigating different design choices when evaluating sensitivity to data injection attacks. Finally, we propose to extend our design algorithm to provide robustness against parameter mismatching.

TABLE III: IO Identification Classification Performance

methods	Horizon	RI	JC	FMI
k-means	1	0.3136	0.0476	0.2054
	2	0.9513	0.2433	0.3916
k-linreg	1	0.70867	0.1123	0.2027
	2	0.9264	0.0342	0.0662

References

[1] S. Tan, J. M. Guerrero, P. Xie, R. Han, and J. C. Vasquez, "Brief survey on attack detection methods for cyber-physical systems," *IEEE Systems Journal*, vol. 14, pp. 5329–5339, Dec 2020.

[2] H. Sandberg, V. Gupta, and K. H. Johansson, "Secure networked control systems," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 445–464, 2022.

[3] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 93–109, 2015.

[4] F. Miao, Q. Zhu, M. Pajic, and G. J. Pappas, "Coding sensor outputs for injection attacks detection," in *53rd IEEE Conf. on Decision and Control*, pp. 5776–5781, 2014.

[5] R. M. G. Ferrari and A. M. H. Teixeira, "A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks," *IEEE Trans. on Automatic Control*, vol. 66, no. 6, pp. 2558–2573, 2021.

[6] P. Griffioen, S. Weerakkody, and B. Sinopoli, "A moving target defense for securing cyber-physical systems," *IEEE Trans. on Automatic Control*, vol. 66, no. 5, pp. 2016–2031, 2021.

[7] H. Guo, Z.-H. Pang, J. Sun, and J. Li, "An output-coding-based detection scheme against replay attacks in cyber-physical systems," *IEEE Trans. on Circuits and Systems II: Express Briefs*, vol. 68, no. 10, pp. 3306–3310, 2021.

[8] M. Ghaderi, K. Gheitasi, and W. Lucia, "A blended active detection strategy for false data injection attacks in cyber-physical systems," *IEEE Trans. on Control of Network Systems*, vol. 8, no. 1, pp. 168–176, 2021.

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

[10] L. Zhai, K. G. Vamvoudakis, and J. Hugues, "Switching watermarking-based detection scheme against replay attacks," in *2021 60th IEEE Conf. on Decision and Control (CDC)*, pp. 4200–4205, 2021.

[11] A. J. Gallo and R. M. G. Ferrari, "Cryptographic switching functions for multiplicative watermarking in cyber-physical systems."

[12] G. Bottegal, F. Farokhi, and I. Shames, "Preserving privacy of finite impulse response systems," *IEEE Control Systems Letters*, vol. 1, no. 1, pp. 128–133, 2017.

[13] V. Katewa, A. Chakrabortty, and V. Gupta, "Differential privacy for network identification," *IEEE Trans. on Control of Network Systems*, vol. 7, no. 1, pp. 266–277, 2020.

[14] P. Stobbe, T. Keijzer, and R. M. Ferrari, "A fully homomorphic encryption scheme for real-time safe control," in *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 2911–2916, IEEE, 2022.

[15] J. Zhang, A. J. Gallo, and R. M. G. Ferrari, "Hybrid design of multiplicative watermarking for defense against malicious parameter identification," 2023.

[16] R. M. Ferrari and A. M. Teixeira, "Detection and isolation of replay attacks through sensor watermarking," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 7363–7368, 2017.

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

[18] J. Katz and Y. Lindell, *Introduction to modern cryptography*. CRC press, 2020.

[19] F. Lauer, G. Bloch, F. Lauer, and G. Bloch, *Hybrid system identification*. Springer, 2019.

[20] L. Bako, G. Mercère, R. Vidal, and S. Lecoeuche, "Identification of switched linear state space models without minimum dwell time," *IFAC Proceedings Volumes*, vol. 42, no. 10, pp. 569–574, 2009.

[21] M. G. Sefidmazgi, M. M. Kordmahalleh, A. Homaifar, and A. Karimoddini, "Switched linear system identification based on bounded-switching clustering," in *2015 American Control Conf. (ACC)*, pp. 1806–1811, 2015.

[22] R. V. Lopes, G. A. Borges, and J. Y. Ishihara, "New algorithm for identification of discrete-time switched linear systems," in *2013 American Control Conf.*, pp. 6219–6224, 2013.

[23] B. Mu, T. Chen, C. Cheng, and E.-w. Bai, "Persistence of excitation for identifying switched linear systems," *Automatica*, vol. 137, p. 110142, 2022.

[24] S. Paoletti, J. Roll, A. Garulli, and A. Vicino, "On the input-output representation of piecewise affine state space models," *IEEE Trans. on Automatic Control*, vol. 55, no. 1, pp. 60–73, 2010.

[25] F. Lauer, "Estimating the probability of success of a simple algorithm for switched linear regression," *Nonlinear Analysis: Hybrid Systems*, vol. 8, pp. 31–47, 2013.