Design of multiplicative watermarking against covert attacks

Gallo, A.J.; Anand, S.C.; Teixeira, Andre M. H.; Ferrari, Riccardo M.G.

DOI
10.1109/CDC45484.2021.9683075

Publication date
2021

Document Version
Final published version

Published in
Proceedings of the 60th IEEE Conference on Decision and Control (CDC 2021)

2021 60th IEEE Conference on Decision and Control (CDC)
December 13-15, 2021. Austin, Texas

Design of multiplicative watermarking against covert attacks

Alexander J. Gallo, Sribalaji C. Anand, André M. H. Teixeira, Riccardo M. G. Ferrari

2021 60th IEEE Conference on Decision and Control (CDC) | 978-1-6654-3659-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CDC45484.2021.9683075

***Abstract*— This paper addresses the design of an active cyber-attack detection architecture based on multiplicative watermarking, allowing for detection of covert attacks. We propose an optimal design problem, relying on the so-called output-to-output ℓ_2-gain, which characterizes the maximum gain between the residual output of a detection scheme and some performance output. Although optimal, this control problem is non-convex. Hence, we propose an algorithm to design the watermarking filters by solving the problem suboptimally via LMIs. We show that, against covert attacks, the output-to-output ℓ_2-gain is unbounded without watermarking, and we provide a sufficient condition for boundedness in the presence of watermarks.**

I. Introduction

Modern engineering systems have been characterized by an ever-growing penetration of cyber resources within physical systems, embedding sensing, communication and computational capabilities due to the reduction of costs of enabling technologies. The scale of the integration has lead to the study of so-called cyber-physical systems (CPS) [1], [2].

Many of the systems that can be appropriately described as CPSs, such as transportation networks, electrical power grids, water distribution networks, among others, are safety critical: indeed, malfunctions in their operation may lead to lack of safety to operators or the general public, as well as economic and societal costs. Apart from accidental malfunctions, given the integration of cyber resources in CPS, these systems have been made the target of malicious attacks, as some high-profile cases show [3], [4], [5].

This has lead to the development of secure control. Differently from the research on cyber-security in information technology, secure control relies on system-theoretic approaches to protect system confidentiality, integrity and availability, and it exploits knowledge of the physical plants at the core of CPS to detect attacks. Control-theoretic cyber security has focused predominantly on cyber-attack detection and resilience, i.e. automatic methods to guarantee a certain level of performance against malicious interference; in this paper we focus on methods for cyber-attack detection. Among the methods that have been proposed in literature, a distinction may be drawn between *passive* and *active* detection methods, where the prior exploit measurements and knowledge of the system to detect the presence of malicious agents, while the latter actively perturb signals to improve the detectability of attacks. While a detailed overview of active methods is out of the scope of this paper, a few examples can be found in [6], [7], [8], [9]. Here we focus on the active detection method proposed in [8], called *multiplicative watermarking*.

Presented in [8] and inspired by authentication schemes with weak cryptographic guarantees, multiplicative watermarking relies on a pair of linear systems, a watermark generator and remover, to modulate the information transmitted between the plant and the controller. This allows for detection of a number of attacks without degrading the performance of the closed-loop CPS. Indeed, the watermark generator and remover are specifically designed such that, the effect of the watermark is removed from the input and output data transmitted between the controller and the plant within the CPS. This improves the detection capabilities of the diagnostic tools of the CPS, as was shown in [8], where the properties of multiplicative watermarking on the output-side communication network of a CPS were investigated.

In this paper we present a method for (sub-)optimal design of multiplicative watermarking units in CPS. Specifically, considering watermarking on both the input and the output-side communication between plant and controller, and relying on output-to-output ℓ_2-gain (OOG) [10], we propose an algorithm minimizing the OOG in the presence of covert attacks. The OOG is an index of worst case gain between the residual input of the diagnosis architecture and a performance output of the plant; as such, it is well suited to be considered as an index for optimal design of control parameters for cyber-attack detection [11], [12]. The contributions of this paper are the following:

a. the formulation of the design of watermarking systems based on OOG;
b. the analysis of OOG against a covert attack [13];
c. the definition of a sufficient condition determining when multiplicative watermarking improves the OOG of the closed-loop system;
d. the design of the watermarking filters to minimize the OOG against covert attacks.

The remainder of the paper is structured as follows: in Section II we present the structure of the CPS with watermarking units, define the attack strategy, and formally introduce the problem. Following this, in Section III, we present the OOG together with some of its fundamental properties. Thus, in Section IV, we show how the OOG may

This work has been partially supported by the Research Council of Norway through the project AIMWind, grant id 312486. This work is supported by the Swedish Research Council under the grant 2018-04396 and by the Swedish Foundation for Strategic Research.

S. C. Anand is with the Department of Electrical Engineering, Uppsala University, PO Box 65, SE-75103, Uppsala, Sweden. (email: sribalaji.anand@angstrom.uu.se

A. M. H. Teixeira is with the Department of Information Technology, Uppsala University, PO Box 337, SE-75105, Uppsala, Sweden. (email: andre.teixeira@it.uu.se

A. J. Gallo and R. M. G. Ferrari are with the Delft Center for Systems and Control, Mechanical, Maritime, and Materials Engineering, TU Delft, Delft, Netherlands (email:a.j.gallo,r.ferrari@tudelft.nl)

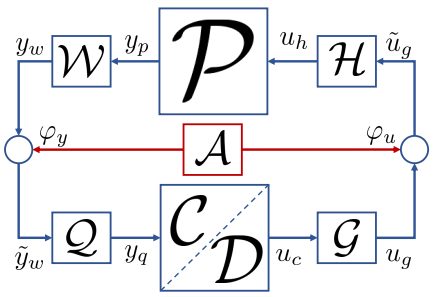

Fig. 1: Block diagram of the closed-loop cyber-physical system, including plant $\mathcal{P}$, controller $\mathcal{C}$ and watermarking filters $\{\mathcal{W}, \mathcal{Q}\}$ and $\{\mathcal{G}, \mathcal{H}\}$. Signals transmitted between the plant and the controller may being altered by an attacker $\mathcal{A}$.

be used for optimal control design, introducing an algorithm to compute the watermarking filters suboptimally, and some of its properties in Sections V and VI. Finally, in Section VII we give a numerical example.

Notation

Let $a : \mathbb{N}_+ \to \mathbb{R}^n$ be a real-valued discrete-time sequence. Given a time horizon $[0, N] \doteq \{k \in \mathbb{N}_+ : 0 \leq k \leq N\}$, the ℓ_2-norm of a over $[0, N]$ is defined as: $\|a\|^2_{\ell_2,[0,N]} \doteq \sum_{k=0}^{N} a[k]^\top a[k]$. Define $\ell_2 \doteq \{x : \mathbb{N}_+ \to \mathbb{R}^n : \|x\|^2_{\ell_2} \doteq \|x\|^2_{\ell_2,[0,\infty]} < \infty\}$ and the extended ℓ_2 space $\ell_{2e} \doteq \{x : \mathbb{N}_+ \to \mathbb{R}^n : \|x\|^2_{\ell_2,[0,N]} < \infty, \forall N \in \mathbb{N}_+\}$.

II. Preliminaries and Problem Formulation

A. System description

We consider a linear time-invariant (LTI) CPS as that shown in Fig. 1, composed of a plant, $\mathcal{P}$, controlled by a dynamic controller $\mathcal{C}$. The control input and measurement output of the plant are transmitted between the plant and the controller over a communication network, and they are modulated through multiplicative watermarking systems [8].

B. Plant and controller

We start the analysis of the closed-loop CPS by defining the physical plant and its controller. The physical plant $\mathcal{P}$ is modeled as a discrete-time LTI system:

$$\mathcal{P} : \begin{cases} x_p^+ = A_p x_p + B_p u_h \\ y_J = C_J x_p + D_J u_h \\ y_p = C_p x_p \end{cases} \tag{1}$$

where $x_p \in \mathbb{R}^n$, $u_h \in \mathbb{R}^m$ are the plant's state and its input, which has been transmitted over the communication network and from which the watermark has been removed by $\mathcal{H}$. The signal $y_p \in \mathbb{R}^p$ is the measured output of the system, while the performance of the system is evaluated over an interval $[0, N], N \in \mathbb{N}$, according to the cost function [14]:

$$J(x_p, u_h) = \|C_J x_p + D_J u_h\|^2_{\ell_2,[0,N]} = \|y_J\|^2_{\ell_2,[0,N]} \tag{2}$$

where $y_J \in \mathbb{R}^{p_J}$ is the virtual performance output of $\mathcal{P}$. All matrices are supposed to be of the correct dimensions.

Assumption 1: We assume that $\mathcal{P}$ is such that (A_p, B_p) is controllable and (C_p, A_p) is observable. ◁

The controller $\mathcal{C}$ is defined as the following:

$$\mathcal{C} : \begin{cases} \hat{x}_p^+ = A_p \hat{x}_p + B_c u_c + L y_r \\ u_c = K \hat{x}_p \\ \hat{y}_p = C_p \hat{x}_p \\ y_r = y_q - \hat{y}_p \end{cases} \tag{3}$$

where $\hat{x}_p \in \mathbb{R}^n$ is an estimate of x_p, $u_c \in \mathbb{R}^m$ is the controller-defined input to the system, $\hat{y}_p \in \mathbb{R}^p$ is the output estimate, and $y_r \in \mathbb{R}^p$ is a residual output which may be used for cyber-attack detection; K and L are chosen to optimize the closed-loop performance.

Remark 1: Definition of the controller in (3) assumes that the controller and the detector are colocated, and thus information available to $\mathcal{C}$ may be used for detection. ◁

We note that, in (1) and (3), we have exploited control input u_h and measurement y_q, rather than u and y_p. These represent the output of the watermark remover systems $\mathcal{H}$ and $\mathcal{Q}$, respectively. As shown in the following, $\mathcal{H}$ and $\mathcal{Q}$ are such that in nominal conditions $u_h = u_c$ and $y_q = y_p$.

C. Watermark generation and removal

The watermark generators and removers are taken to be linear systems, for which series interconnection is the identity. Specifically, the input and output watermarking pairs are defined as $\mathcal{G}, \mathcal{H}$ and $\mathcal{W}, \mathcal{Q}$, respectively. Their dynamics are given by the following:

$$\Sigma : \begin{cases} x_\sigma^+ = A_\sigma x_\sigma + B_\sigma \nu_\sigma \\ \gamma_\sigma = C_\sigma x_\sigma + D_\sigma \nu_\sigma \end{cases} \tag{4}$$

where subscript $\sigma \in \{g, h, w, q\}$ defines whether the state, input, or output are associated with $\mathcal{G}, \mathcal{H}, \mathcal{W}, \mathcal{Q}$. The systems are square, and

$$\begin{aligned} \left[\nu_g^\top, \ \nu_h^\top, \ \nu_w^\top, \ \nu_q^\top\right]^\top &\doteq \left[u_c^\top, \ \tilde{u}_g^\top, \ y_p^\top, \ \tilde{y}_w^\top\right]^\top, \\ \left[\gamma_g^\top, \ \gamma_h^\top, \ \gamma_w^\top, \ \gamma_q^\top\right]^\top &\doteq \left[u_g^\top, \ u_h^\top, \ y_w^\top, \ y_q^\top\right]^\top, \end{aligned} \tag{5}$$

where a tilde is added to a variable to highlight it as being transmitted over a communication network, and therefore possibly subject to attack, as per Fig. 1. All matrices are of appropriate dimensions, all systems are stable.

Remark 2: Given that $\mathcal{G}, \mathcal{H}, \mathcal{W}, \mathcal{Q}$ are defined by the system operator, their stability can be guaranteed. ◁

The watermarking pairs are designed as in [8], i.e.:

$$\mathcal{H} \doteq \mathcal{G}^{-1} \quad \mathcal{Q} \doteq \mathcal{W}^{-1}, \tag{6}$$

where the inverse of a system is given in Definition 1.

Definition 1 ([14, Lemma 3.15]): Define the transfer function from the tuple (A, B, C, D) as:

$$G(z) = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right], \tag{7}$$

and suppose that D is an invertible matrix. Then

$$G^{-1}(z) = \left[\begin{array}{c|c} A - BD^{-1}C & -BD^{-1} \\ \hline D^{-1}C & D^{-1} \end{array}\right] \tag{8}$$

is the inverse transfer function of $G(z)$. $\triangleleft$

Given this definition, in the absence of attacks, we have:

$$\mathcal{Q}(z)\mathcal{W}(z) = I_p\,, \quad \mathcal{H}(z)\mathcal{G}(z) = I_m\,, \tag{9}$$

$$u_h[k] = u[k]\,, \quad y_q[k] = y_p[k]\,, \quad \forall k \in \mathbb{N}_+, \tag{10}$$

assuming $x_s[0] = x_t[0], s \in \{g, w\}, t \in \{h, q\}$.

D. Cyber-attack modeling

We consider a malicious agent $\mathcal{A}$, as in Fig. 1, capable of injecting attacks to the signals transmitted between the controller and the plant. This is formally modeled as:

$$\tilde{u}_g[k] \doteq u_g[k] + \beta_u[k - K_a^u]\varphi_u[k] \tag{11a}$$

$$\tilde{y}_w[k] \doteq y_w[k] + \beta_y[k - K_a^y]\varphi_y[k] \tag{11b}$$

where $\varphi_u[k]$ and $\varphi_y[k]$ are actuator and sensor attack sequences defined by the malicious agent. For $l \in \{u, y\}$, the function $\beta_l[\cdot]$ is an activation function, defined as:

$$\beta_l[k] = \begin{cases} (1 - b_l^k)\,, & \text{if } k \geq 0 \\ 0\,, & \text{otherwise} \end{cases}, \tag{12}$$

and where $K_a^l > 0$ is the initial instant of attack, and $b_l \in [0, 1]$; without loss of generality, we assume $b_l = 0$.

We assume that $\mathcal{A}$ has the necessary resources (as defined in [15]) to leverage a *covert attack* against the CPS without watermarking, i.e. that φ_u and φ_y satisfy:

$$\mathcal{A} : \begin{cases} x_a^+ = A_p x_a + B_p \varphi_u \\ y_a = C_p x_a \\ \varphi_y = -y_a \end{cases} \tag{13}$$

with internal state $x_a \in \mathbb{R}^n$, and where φ_u is arbitrarily defined by the attacker, such that $\varphi_u \in \ell_{2e}$, and with $K_a^u = K_a^y = K_a$. We consider that the malicious agent does not have knowledge of the watermarking systems $\{\mathcal{W}, \mathcal{Q}, \mathcal{G}, \mathcal{H}\}$.

E. Cyber-attack detection

Given the possibility of cyber-attacks, we equip the controller with detection logic. We use the innovation y_r as a *residual*, compared to an appropriately defined threshold θ_r, designed to satisfy the trade-off between ability of detecting attacks and robustness against noise. For the purpose of this paper, the threshold is set as $\theta_r = 1$. Thus, for $N > 0$, the detection test may be formalized as:

$$\|y_r\|^2_{\ell_2,[0,N]} \geq \theta_r. \tag{14}$$

It is known that, in the absence of watermarking, it is sufficient for the attacker to select $x_a[K_a] = 0$ for the attack defined in (11)-(13) to be stealthy, i.e. for $\varphi \doteq \left[\varphi_u^\top, \varphi_y^\top\right]^\top$ to not influence the residual output y_r [13], [16].

F. Problem formulation

Having presented the overall architecture for cyber-attack detection with watermarking, we can formally present the objective of this paper. Let us recall the system's OOG [10]: introduced as a metric to quantify the effect of worst-case stealthy attacks on the performance of a system, it measures the amplification between the residual and performance outputs, y_r and y_p, respectively. We introduce the following:

$$\mathcal{S} : \begin{cases} x^+ = Ax + Ba \\ y_1 = C_1 x + D_1 a \\ y_2 = C_2 x + D_2 a \end{cases} \tag{15}$$

where $x \in \mathbb{R}^\nu$ is the closed loop system state, $a \in \mathbb{R}^\mu$ is the attack signal, $y_1 \in R^{\upsilon_1}$ is a residual output used for detection and $y_2 \in \mathbb{R}^{\upsilon_2}$ is the performance output. The system $\mathcal{S}$ can be seen as any closed loop system $\mathcal{P}, \mathcal{C}$ driven by an external attack signal, with y_r, y_J and φ substituted with y_1, y_2 and a, respectively.

Definition 2: Take $\mathcal{S}$ as in (15), the output-to-output ℓ_2-gain is defined as:

$$\|\mathcal{S}\|^2_{\ell_2, y_2 \leftarrow y_1} \doteq \begin{array}{ll} \sup\limits_{a \in \ell_{2e}} & \|y_2\|^2_{\ell_2} \\ s.t. & \|y_1\|^2_{\ell_2} \leq 1, \quad x[0] = 0 \end{array}. \tag{16}$$

$\triangleleft$

We now formulate the central problem of this paper.

Problem 1: Design the parameters of $\{\mathcal{W}, \mathcal{Q}, \mathcal{G}, \mathcal{H}\}$ such that (9) holds, while minimizing the system OOG. $\triangleleft$

III. Output-to-output ℓ_2-gain

Let us briefly summarize the main results in [10], to introduce the main properties of the output-to-output ℓ_2-gain.

As shown in [10], the non-convex optimization problem (16), can be cast into its convex dual

$$\|\mathcal{S}\|^2_{\ell_2, y_2 \leftarrow y_1} \doteq \begin{array}{ll} \inf\limits_{\gamma > 0} & \gamma \\ s.t. & \|y_2\|^2_{\ell_2} \leq \gamma \|y_1\|^2_{\ell_2}\,, \quad \forall a \in \ell_{2e} \\ & x[0] = 0 \end{array}. \tag{17}$$

Thus, recalling Definition 2, and given that y_1 is taken to be a suitable residual output of $\mathcal{S}$, defining $\gamma^* \doteq \|\mathcal{S}\|^2_{\ell_2, y_2 \leftarrow y_1}$, note that it can be interpreted as the maximum amplification of the system from $\|y_1\|^2_{\ell_2}$ to $\|y_2\|^2_{\ell_2}$, i.e.

$$\|y_2\|^2_{\ell_2} \leq \gamma^* \|y_1\|^2_{\ell_2}\,, \quad x[0] = 0. \tag{18}$$

Furthermore, recalling (16), γ^* also represents the worst case impact of an attack on the performance of the system:

$$\|y_2\|^2_{\ell_2} \leq \gamma^* \|y_1\|^2_{\ell_2} \leq \gamma^*\,, \quad \forall a \in \ell_{2e} \tag{19}$$

Following terminology in [10], any attack capable of achieving the worst case gain γ^*, the optimal solution of (17), is said to be a *strategic attack*.

Although more readily solvable than (16), the optimization problem (17) is formulated in signal space, and is therefore infinite dimensional. By relying on results from dissipative system theory, it can be shown[1] that the following holds.

[1]We refer the interested reader to [11] for details.

Proposition 1: [11, Prop.1] Consider the LTI system $\mathcal{S}$ defined in (15), and presume that (A, B) is controllable and (C_1, A) is observable. Define a supply function

$$s(x, a) = \gamma \|y_1[k]\|_2^2 - \|y_2[k]\|_2^2 \tag{20}$$

Then the following statements are equivalent:

a. The system $\mathcal{S}$ is dissipative w.r.t. $s(x, a)$;

b. For all trajectories of x, and $N > 0$ and $x[0] = 0$,

$$\sum_{\kappa=0}^{N-1} s(x[\kappa], a[\kappa]) \geq 0. \tag{21}$$

c. There exists some $P \succcurlyeq 0$ such that

$$R(P) - \gamma \begin{bmatrix} C_1^\top \\ D_1^\top \end{bmatrix} \begin{bmatrix} C_1 \, D_1 \end{bmatrix} + \begin{bmatrix} C_2^\top \\ D_2^\top \end{bmatrix} \begin{bmatrix} C_2 \, D_2 \end{bmatrix} \preccurlyeq 0. \tag{22}$$

with $R(P)$ defined as:

$$R(P) = \begin{bmatrix} A^\top P A - P & A^\top P B \\ A^\top P B & B^\top P B \end{bmatrix}. \tag{23}$$

□

Thus, considering supply function $s(x, a)$ defined in (20), in light of Proposition 1, and as shown in [10], it is possible to compute the OOG of $\mathcal{S}$ as $\gamma^* = \|\mathcal{S}\|^2_{\ell_2, y_2 \leftarrow y_1}$, with

$$\begin{aligned} \gamma^* = \min_{P,\gamma} \quad & \gamma \\ s.t. \quad & P \succcurlyeq 0 \, , \gamma > 0 \\ & R(P) - \gamma \begin{bmatrix} C_1^\top \\ D_1^\top \end{bmatrix} \begin{bmatrix} C_1 \, D_1 \end{bmatrix} + \begin{bmatrix} C_2^\top \\ D_2^\top \end{bmatrix} \begin{bmatrix} C_2 \, D_2 \end{bmatrix} \preccurlyeq 0 \end{aligned} . \tag{24}$$

Proposition 2 ([10, Th. 2]): Consider the LTI system $\mathcal{S}$ defined in (15) with OOG γ^*. The OOG is finite if and only if either of the following conditions hold:

a. the system (A, B, C_1, D_1) has no unstable zeros associated with reachable x_λ;

b. the unstable zeros of (A, B, C_1, D_1) associated with a reachable x_λ are also zeros of (A, B, C_2, D_2);

where $x_\lambda \neq 0$ is the eigenvector associated to $\lambda \in \sigma(A)$. □

As pointed out in [11], the OOG has a fundamental limitation, summarized in the following proposition.

Proposition 3 ([11, Lem. 1]): Let $D_2 \neq 0$, full column rank, and $D_1 = 0$. Then the OOG of $\mathcal{S}$ is unbounded. □

In light of Proposition 3, in the remainder of this article, we assume that $D_2 = 0$, i.e. that $D_J = 0$ in (1).

Remark 3: A consequence of assuming $D_2 = 0$ is that the performance cost $J(x, u)$ is evaluated as a function of the state alone. Although this may be restrictive in general, as it does not pose any cost on the energy required for actuation, it permits the use of OOG defined in (16) rather than the *truncated* OOG presented in [11]. Analysis of optimality with the latter is left as future work. These limitations can be remedied by augmenting the plant dynamics with actuator dynamics, or by introducing a time-delay in the application of the input, as in [12]. ◁

Finally, before designing the controller matrices, it is important to note that the formulation of the OOG implicitly presumes that the attacker has full knowledge of the closed loop dynamics $\mathcal{S}$. As will be shown in the following, this will be fundamental when dealing with the performance of the closed-loop dynamics including watermarking.

IV. Optimal controller design

Having presented some background information on the OOG, we are now ready to discuss the design of optimal control gains K and L in $\mathcal{C}$ under covert attacks. Consider the closed-loop system in (15) for the system defined by the feedback interconnection of $\mathcal{P}$ and $\mathcal{C}$, with state $x \doteq \left[x_p^\top, \, e^\top\right]^\top$, with $e \doteq x_p - \hat{x}_p$, $a \doteq \varphi$, and $y \doteq \left[y_r^\top, \, y_J^\top\right]^\top$. The system matrices can be derived from the feedback interconnection of $\mathcal{P}$ and $\mathcal{C}$. Clearly, the closed loop dynamics of the system depend on the definition of K and L in $\mathcal{C}$. Thus, they can be included as decision variables in the optimization problem defining $\|\mathcal{S}\|_{\ell_2, y_2 \leftarrow y_1}$:

$$\begin{aligned} \min_{P,\gamma,K,L} \quad & \gamma \\ \text{s.t.} \quad & P \succcurlyeq 0 \, , \, \gamma > 0 \, , \\ & \begin{bmatrix} A^\top P A - P & A^\top P B \\ B^\top P A & B^\top P B \end{bmatrix} - \gamma \begin{bmatrix} C_1^\top \\ D_1^\top \end{bmatrix} \begin{bmatrix} C_1 & D_1 \end{bmatrix} \\ & + \begin{bmatrix} C_2 & D_2 \end{bmatrix}^\top \begin{bmatrix} C_2 & D_2 \end{bmatrix} \preccurlyeq 0 \end{aligned} \tag{25}$$

Theorem 1: Consider $\mathcal{S}$ subject to covert attack (13). The OOG is unbounded, irrespective of K and L. □

Proof: In the interest of space, we omit the proofs. ■

Theorem 1 shows that if the malicious agent has the capabilities of performing a covert attack, the OOG is not a suitable criterion for the design of the control matrices. Therefore, other approaches may be preferred for its design.

V. Watermarking system design

Having shown that, in the absence of watermarking systems, there are no control gains K and L such that $\|\mathcal{S}\|^2_{\ell_2, y_J \leftarrow y_r}$ is bounded, we now show how including watermarking units (4) may be used to improve the closed-loop performance against covert attacks.

To this end, take the closed-loop system $\mathcal{S}$ in (15) to represent the closed-loop CPS composed of the feedback interconnection of $\mathcal{P}$ and $\mathcal{C}$ together with the watermarking units $\{\mathcal{W}, \mathcal{Q}\}$ and $\{\mathcal{G}, \mathcal{H}\}$, as shown in Fig. 1. Thus, defining $x \doteq \left[x_p^\top, \, e^\top, \, x_g^\top, \, x_h^\top, \, x_w^\top, \, x_q^\top, \, x_a^\top\right]^\top$, $a \doteq \varphi$, and $y \doteq \left[y_r^\top, \, y_J^\top\right]^\top$, the closed loop dynamics are described by (15) with matrices (A, B, C, D) found by taking the feedback connection of the plant $\mathcal{P}$, the output-side watermarking systems $\{\mathcal{W}, \mathcal{Q}\}$, the controller $\mathcal{C}$, and the input-side watermarking pair $\{\mathcal{H}, \mathcal{G}\}$, as shown in Fig. 1. For ease of notation, let us define $\mathbf{W} \doteq \{\mathcal{W}, \mathcal{Q}, \mathcal{G}, \mathcal{H}\}$.

In light of the discussion of the OOG in the previous section, it is possible to formulate the following optimization problem, setting the watermarking system parameters as decision variables and minimizing the OOG $\|\mathcal{S}\|^2_{\ell_2, y_2 \leftarrow y_1}$:

$$\begin{aligned} \min_{P,\gamma,\mathbf{W}} \quad & \gamma \\ \text{s.t.} \quad & P \succcurlyeq 0 \, , \, \gamma > 0 \\ & R(P) - \gamma \left[C_1, \, D_1\right]^\top \left[C_1, \, D_1\right] \\ & \quad + \left[C_2, \, D_2\right]^\top \left[C_2, \, D_2\right] \preccurlyeq 0. \end{aligned} \tag{26}$$

with $R(P)$ defined in (23), and C_1, C_2, D_1, and D_2 are such that $C = \begin{bmatrix} C_1^\top, C_2^\top \end{bmatrix}^\top$ and $D = \begin{bmatrix} D_1^\top, D_2^\top \end{bmatrix}^\top$. Because of the definition of the closed loop matrices (A, B, C, D) with respect to the watermarking parameters, as well as the definition of $R(P)$, problem (26) is non-convex. Thus, to solve it *suboptimally* via an LMI formulation, we consider an alternating minimization algorithm, in which the solution is found iteratively by solving for P while fixing the other decision variables, and then solving for the parameters of $\mathcal{S}$ with the value of P fixed, until a stopping criterion is met [17]. This leads to LMI constraints when solving for P, although not when solving for the parameters of $\mathcal{S}$.

To avoid this, we consider that the watermark systems have some predefined *structure*, i.e. that some of the matrices defining them are decided *a priori*, to linearize the constraints of (26) when P is fixed. A number of different approaches may be taken, such as defining finite impulse or infinite impulse response (FIR and IIR) filters for each of the components of u_c and y_m, or defining D_s and C_s *a priori* as $\bar{D}_s$ and $\bar{C}_s$, respectively, for $s \in \{h, q\}$.

Thus, the optimization problem (26) is redefined, taking the Schur complement of $R(P)$:

$$\min_{P,\gamma,\mathbf{W},P_q,P_h} \gamma \tag{27a}$$

$$\text{s.t.} \quad P \succcurlyeq 0\,,\ \gamma > 0\,,\ P_q \succcurlyeq 0\,,\ P_h \succcurlyeq 0 \tag{27b}$$

$$\begin{bmatrix} -P_s & A_s^\top P_s \\ P_s A_s & -P_s \end{bmatrix} \preccurlyeq 0 \quad s \in \{h, q\} \tag{27c}$$

$$A_w = A_q - B_q D_q^{-1} C_q\,, \tag{27d}$$

$$A_g = A_h - B_h D_h^{-1} C_h\,, \tag{27e}$$

$$\begin{bmatrix} -P & 0 & A^\top P \\ 0 & 0 & B^\top P \\ PA & PB & -P \end{bmatrix} - \gamma \begin{bmatrix} C_1^\top \\ D_1^\top \\ 0 \end{bmatrix} \begin{bmatrix} C_1 & D_1 & 0 \end{bmatrix} + \begin{bmatrix} C_2^\top \\ D_2^\top \\ 0 \end{bmatrix} \begin{bmatrix} C_2 & D_2 & 0 \end{bmatrix} \preccurlyeq 0, \tag{27f}$$

$$C_s = \bar{C}_s, \quad D_s = \bar{D}_s, \quad s \in \{h, q\}, \tag{27g}$$

where (27c) is included, together with positive semidefiniteness of $P_s, s \in \{h, q\}$, to guarantee the stability of the watermarking systems, and (27d)–(27e) guarantee that (9) hold, and therefore that $\{\mathcal{G}, \mathcal{H}\}$ and $\{\mathcal{W}, \mathcal{Q}\}$ are indeed watermarking pairs. Notice that (27) is bilinear in the constraints. Algorithm 1 solves the problem suboptimally, by implementing an alternating algorithm [17].

Remark 4: Note that, to achieve the worst case gain γ^*, the attacker must have full knowledge of the watermarking systems' parameters. ◁

VI. Structural constraints on solvability of Algorithm 1

Let us briefly comment on the feasibility of Algorithm 1, given the covert attack scenario considered in (11) with covert attack defined in (13). If the algorithm is infeasible, this implies that there exists an input sequence $\varphi_u \in \ell_{2e}$ such that the covert attack strategy defined in (13) remains undetectable in the presence of the watermarking units.

Algorithm 1 Watermark design

Input: Stabilizing K, L, stable $\mathbf{W}$, $\epsilon \in \mathbb{R}_+$
Output: $\mathbf{W}^*, \gamma^*, P^*$

Set $k = 0$, $\gamma_0 = 0$, $\gamma_{-1} = \infty$, $\mathbf{W}_0 = \{\mathcal{W}, \mathcal{Q}, \mathcal{H}, \mathcal{G}\}$
while $\|\gamma_k - \gamma_{k-1}\| > \epsilon$ **do**
    Find P_{k+1} optimizing (27) w.r.t. P with $\mathbf{W} = \mathbf{W}_k$;
    Find $\mathbf{W}_{k+1}, , \gamma_{k+1}$ optimizing (27) w.r.t. γ, $\mathbf{W}$, P_q, P_h with $P = P_{k+1}$;
    $k = k + 1$;
end while
return: $\mathbf{W}^* = \mathbf{W}_k, \gamma^* = \gamma_k, P^* = P_k$.

Given the structure of $\mathcal{C}$, an attack that is undetectable will also guarantee $y_q^a = 0$, where y_q^a is the component of y_q driven by the attack φ_u. In turn, given invertibility of $\mathcal{Q}$, it is possible to see that, if φ_u is an undetectable sequence, then $\tilde{y}_w^a \doteq y_w^a - y_a = 0$ for all $k \geq K_a$, where, again, superscript a is used to define the component of y_w driven by the attack input φ_u. Formally, $\tilde{y}_w^a$ can be seen as the output of $\mathcal{S}^a$:

$$\mathcal{S}^a : \begin{cases} x^{a+} = A^a x^a + B^a \varphi_u \\ y^a = C^a x^a + D^a \varphi_u \end{cases} \tag{28}$$

where $x^a \doteq \begin{bmatrix} x_h^\top, x_p^\top, x_w^\top, x_a^\top \end{bmatrix}^\top$, $y^a \doteq \tilde{y}_w^a$, and with matrices defined considering the series connection of $\mathcal{H}, \mathcal{P}, \mathcal{W}$ and the attacker-defined system $\mathcal{A}$ (13).

It is well known that an attack $\varphi_u \neq 0$ against $\mathcal{S}^a$ is undetectable, and therefore is such that $y^a = 0$, if and only if it is a zero-dynamics attack, i.e. it satisfies:

$$\begin{bmatrix} \lambda I - A^a & -B^a \\ C^a & D^a \end{bmatrix} \begin{bmatrix} \bar{x}^a \\ \bar{\varphi}_u \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{29}$$

for some $\lambda \in \mathbb{C}$ and some $\bar{x}^a$ and $\bar{\varphi}_u$ [18]. $G_w(z), G_p(z), G_h(z), G_a(z)$ are the transfer functions of $\mathcal{W}, \mathcal{P}, \mathcal{H}$ and $\mathcal{A}$, respectively, with $G_a(z) \equiv G_p(z), \forall z \in \mathbb{C}$.

Lemma 1: Consider (28). The following are equivalent:
a. Exists $\lambda \in \mathbb{C}$, $\bar{x}^a \in \mathbb{R}^n$, $\bar{\varphi}_u \in \mathbb{R}^m$: (29) holds;
b. Exists $\mu \in \mathbb{C}$: $G_w(\mu) G_p(\mu) G_h(\mu) \equiv G_p(\mu)$, $\varphi \neq 0$. □

Proof: In the interest of space, we omit the proofs. ■

We now present the main theoretical result of this paper: a sufficient condition under which Algorithm 1 is feasible, guaranteeing the output-to-output gain of the closed-loop system with watermarking, γ, is finite.

Theorem 2: Consider $\mathcal{S}$ in (15). Suppose that $\mathcal{P}$ defined in (1) satisfies condition a. in Proposition 2. Thus, Algorithm 1 returns a solution such that $\gamma < \infty$, so long as the initial choice of $\mathbf{W}$ is such that (27) is feasible. □

Proof: In the interest of space, we omit the proofs. ■

VII. Numerical example

In this section, the effectiveness of the proposed Algorithm 1 is depicted through a numerical example. Consider a plant $\mathcal{P}$ and controller $\mathcal{C}$ with the following parameters: $A_p = \begin{bmatrix} 0.9191 & 0.3277 \\ -0.0768 & 0.4269 \end{bmatrix}, B_p = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, C_J =$

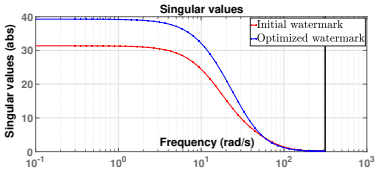

Fig. 2: Singular values for the performance output

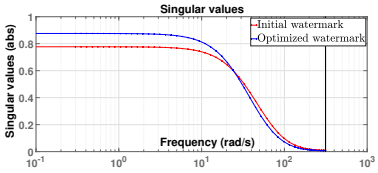

Fig. 3: Singular values for the detection output

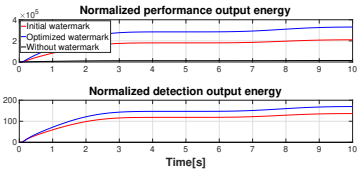

Fig. 4: System outputs for a covert attack

$\begin{bmatrix}2\\0\end{bmatrix}^\top, D_J = 0, C_p = \begin{bmatrix}1\\0\end{bmatrix}^\top, K = \begin{bmatrix}-0.3405\\-0.3987\end{bmatrix}^\top$, and $L = \begin{bmatrix}0.5956\\-0.0253\end{bmatrix}$. We consider a watermark remover at the output and the input with the structure $B_f = C_f = D_f = 1, f \in \{q, h\}$. Note that here we consider a more constrained case than the general case, where either B_f or C_f must be known *a priori*. The watermark generator and remover are related by (27d)–(27e). We initialize the algorithm with $\epsilon = 10^{-5}$ and the stable watermark state-transition matrices $A_q = 0.6714$ and $A_h = 0.5201$. Firstly, the plant $\mathcal{P}$ has no unstable zeros and hence the limitation discussed in Section VI does not apply for the system in consideration. That is, there does not exist an input sequence which is 0-stealthy to the detector. Although, when the watermarking scheme is absent, (27) will be unbounded since the adversary knows the system.

The objective of the OOG is to design the watermark adder and remover such that the gain (or singular values (SVs)) of the system from the attack input to the performance output is decreased, whilst the gain of the system from the attack input to the detection output is increased at all frequencies. To this end we represent the SVs, on the unit circle of the complex plane, of the system from the attack input to performance and detection outputs in Fig. 2 and Fig. 3. It is evident from Fig. 2 that, from $\omega = 0$ rad/s to $\omega \approx 25$ rad/s, the performance of the system deteriorates as the SVs increase on optimizing. But the algorithm compensates for this deterioration by simultaneously increasing the detection performance as can be seen in Fig. 3. The SVs of both the systems do not change much from $\omega \approx 25$ rad/s. This is because, the output-to-output gain focuses on improving the detection performance only when the performance loss is significant and vice versa. This was also pointed out in [12].

Let us now consider an attack signal of the form (13) where $\varphi_u \doteq 5 + 5sin(k)$. The normalized energies of the performance and detection outputs with the initial watermark parameters and the optimized watermark parameters (obtained from Algorithm 1) are shown in Fig. 4. The attack is undetectable in the absence of watermarks. Furthermore, while the effect of the attack on the performance output is increased, the effect of the attack on the detection output is increased by at least 25%.

VIII. Conclusions

In this work we have presented the optimal design of multiplicative watermarking based on the OOG of systems. We show how, by including multiplicative watermarking on the input and output channels of the system, its OOG can be made finite in the presence of covert attacks. As future work, we wish to further study structural conditions under which multiplicative watermarking may bound the OOG of the system in the presence of covert attacks. Further analysis is also required to solve the design procedure optimally.

References

[1] R. Baheti and H. Gill, "Cyber-physical systems," *The impact of control technology*, vol. 12, pp. 161–166, 2011.

[2] J. Giraldo, E. Sarkar, A. A. Cardenas, M. Maniatakos, and M. Kantarcioglu, "Security and privacy in cyber-physical systems: A survey of surveys," *IEEE Design & Test*, vol. 34, no. 4, pp. 7–17, 2017.

[3] N. Falliere, L. O. Murchu, and E. Chien, "W32. stuxnet dossier," *White paper, Symantec Corp., Security Response*, vol. 5, no. 6, p. 29, 2011.

[4] R. M. Lee, M. J. Assante, and T. Conway, "Analysis of the cyber attack on the Ukrainian power grid," *Electricity Information Sharing and Analysis Center (E-ISAC)*, 2016.

[5] B. Sobczak, *Denial of Service attack caused grid cyber disruption: DOE*. Environment & Energy Publishing, 2019.

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

[7] ——, "Challenges and opportunities: Cyber-physical security in the smart grid," in *Smart Grid Control*. Springer, 2019, pp. 257–273.

[8] R. M. Ferrari and A. M. Teixeira, "A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks," *IEEE Trans. on Automat. Contr.*, 2020.

[9] A. J. Gallo, M. S. Turan, F. Boem, G. Ferrari-Trecate, and T. Parisini, "Distributed watermarking for secure control of microgrids under replay attacks," *IFAC-PapersOnLine*, vol. 51, no. 23, pp. 182–187, 2018.

[10] A. Teixeira, H. Sandberg, and K. H. Johansson, "Strategic stealthy attacks: the output-to-output ℓ_2-gain," in *2015 54th IEEE Conf. on Decision and Contr. (CDC)*. IEEE, 2015, pp. 2582–2587.

[11] A. M. Teixeira, "Optimal stealthy attacks on actuators for strictly proper systems," in *2019 IEEE 58th Conf. on Decision and Contr. (CDC)*. IEEE, 2019, pp. 4385–4390.

[12] S. C. Anand and A. Teixeira, "Joint controller and detector design against data injection attacks on actuators," in *IFAC world congress 2020 (Virtual), Berlin, Germany, July 11-17*, 2020.

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

[14] K. Zhou, J. C. Doyle, K. Glover *et al.*, *Robust and optimal control*. Prentice hall New Jersey, 1996, vol. 40.

[15] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135–148, 2015.

[16] A. Barboni, H. Rezaee, F. Boem, and T. Parisini, "Detection of covert cyber-attacks in interconnected systems: a distributed model-based approach," *IEEE Trans. on Automat. Contr.*, vol. 65, no. 9, pp. 3728–3741, 2020.

[17] Q. Li, Z. Zhu, and G. Tang, "Alternating minimizations converge to second-order optimal solutions," in *International Conf. on Machine Learning*. PMLR, 2019, pp. 3935–3943.

[18] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. on Automat. Contr.*, vol. 58, no. 11, pp. 2715–2729, 2013.