This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2020.2971994, IEEE Transactions on Automatic Control

An Online Approach to Physical Watermark Design

Hanxiao Liu, Yilin Mo[†], Jiaqi Yan, Lihua Xie, and Karl H. Johansson

***Abstract*—This paper considers the problem of designing physical watermark signals in order to optimally detect possible replay attack in a linear time-invariant system, under the assumption that the system parameters are unknown and need to be identified online. We first provide a replay attack model, where an adversary replays the previous sensor data in order to fool the system. A physical watermarking scheme, which leverages a random input as a watermark to detect the replay attack, is then introduced. The optimal watermark signal design problem is cast as an optimization problem, which aims to achieve the optimal trade-off between control performance and intrusion detection. An online watermarking design and system identification algorithm is provided to deal with systems with unknown parameters. We prove that the proposed algorithm converges to the optimal one and characterize the almost sure convergence rate. An industrial process example is provided to illustrate the effectiveness of the proposed strategy.**

***Index Terms*—Cyber-Physical System, Security, Intrusion Detection, System Identification**

I. Introduction

Cyber-Physical Systems (CPSs) offer close integration of computational elements and physical processes [1]. Such systems play a critical role in large varieties of fields, such as manufacturing, health care, environment control, transportation, etc. Due to their wide applications and critical functions, it is of paramount importance to ensure the secure operation of CPS [2], [3]. Any successful attack on CPS may jeopardize critical infrastructure and people's lives and properties, even threaten national security. In 2010, Stuxnet malware launched a devastating attack on Iranian uranium enrichment facilities [4]. This incident raised a great deal of attention to CPS security in recent years [5].

A significant amount of research effort has been devoted to intrusion and anomaly detection algorithms to enhance CPS security. Mitchell and Chen [9] proposed a hierarchical performance model and techniques for intrusion detection in CPS. They classified the modern CPS intrusion detection system techniques into two classes: detection technique and audit material. They summarized advantages and disadvantages in [10]. Kwon *et al.* [11] discussed necessary and sufficient conditions under which the attacker could be successful without being detected. Their method can be employed to evaluate vulnerability degree of certain CPSs. Corresponding detection and defense methodologies against stealthy deception attacks can be developed. In [12], the authors proposed a mathematical framework for CPS and investigated limitations of the monitoring system. Centralized and distributed attack detection and identification monitors were also discussed.

In this paper, we consider the detection problem of replay attacks. In [13], [14], [15], a replay attack model is defined and its effect on a steady-state control system is analyzed. An algebraic condition is provided on the detectability of the replay attack. For those systems that cannot detect replay attack efficiently, a physical watermarking scheme is proposed to enable the detection of a replay attack. In particular, by injecting a random control signal, the watermark signal, into the control system, it is possible to secure the system. However, the watermark signal may deteriorate the control performance, and therefore it is important to find the optimal trade-off between the control performance and the detection efficiency, which can be cast as an optimization problem. Similar watermarking scheme is also proposed in the literature [16].

Different from the previous additive watermarking schemes, a multiplicative sensor watermarking scheme is proposed in [17]. In this scheme, each output is respectively fed to a SISO watermark generator and due to the inclusion of a watermark removing functionality, the control performance will not be sacrificed. Applying some techniques of non-cooperative stochastic games, Miao *et al.* [18] designed a suboptimal switching control policy that balances control performance and the intrusion detection rate for replay attacks. Hoehn and Zhang [19] provided a novel technique via exciting the system in non-regular time intervals and signal processing to detect the replay attack.

It is worth noticing that in majority of the aforementioned research, the precise knowledge of the system parameters is required in order to design the watermark signal and the detector. However, acquiring these parameters may be troublesome and costly. Moreover, for a large system, the system parameters may change during its operation. Hence, it is beneficial for the system to learn the parameters in an online fashion and automatically generate the optimal detector and the watermark signal in real-time. The problem of learning parameters of dynamical systems, i.e., system identification, has been studied over the past decades. Most methods, however, require persistent excitation on the input.

The goal of this paper is to develop a data-driven approach to design physical watermark signals to protect systems with unknown parameters, against replay attack. In this paper, due to the nature of the optimal watermark signal, we shall design the input that asymptotically converges to a signal that does not satisfy the persistent excitation condition. However, by controlling the convergence rate, we can still prove that the system parameters converge to the true parameters almost surely.

Some preliminaries results regarding online design of physical watermarks are contained in our former work [20]. The main differences between the current version of the paper and [20] are: 1) we not only prove that we can asymptotically identify the system parameters, but also characterize the rate of the convergence; 2) we provide a procedure to automatically generate the Neyman-Pearson (NP) detector; 3) we provide the simulation on an industrial process to verify the effectiveness of the proposed approach.

[†]: Corresponding Author.

H. Liu is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, and the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden. Email: hanxiao001@ntu.edu.sg.

Y. Mo is with the Department of Automation and BNRist, Tsinghua University, China. Email: ylmo@tsinghua.edu.cn.

J. Yan and L. Xie are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. Email: {jyan004, elhxie}@ntu.edu.sg.

K.H. Johansson is with the Division of Decision and Control Systems, the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden. Email: kallej@kth.se.

This work is supported by the A*STAR Industrial Internet of Things Research Program, under the RIE2020 IAF-PP Grant A1788a0023, the National Key Research and Development Program of China under Grant 2018AAA0101601, the Knut and Alice Wallenberg Foundation, the Swedish Foundation for Strategic Research, and the Swedish Research Council.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2020.2971994, IEEE Transactions on Automatic Control

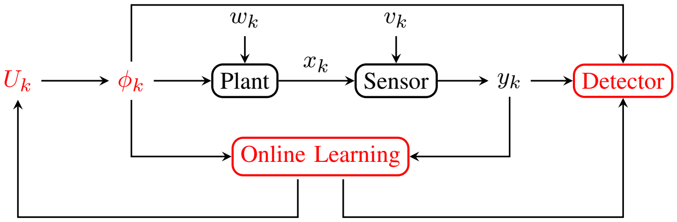

Fig. 1. The system diagram.

The main contributions of this paper are threefold: 1) An online "learning" algorithm is presented to simultaneously infer the parameters of the system based only on the system input and output data and generate the watermark signal as well as the optimal detector based on the estimated parameters. To the best of our knowledge, it is the first time to study the detection of replay attacks under the scenario with unknown system parameters. 2) We prove that the system parameters which are inferred via our proposed online algorithm converge to the true parameters almost surely even if the input signal asymptotically converges to a degenerate signal. 3) We also characterize the almost sure convergence rate of the estimated system parameters to the true parameters and provide an upper bound for this rate.

Notations: $\|A\|$ of the matrix A is the spectral norm of an $m \times n$ matrix A, which is its largest singular value. $A \otimes B$ is the Kronecker product of matrices A and B. $A > 0$ ($A \geq 0$) indicates that A is positive definite (positive semidefinite). A^+ denotes the pseudo-inverse of A. We say that $f(k) \sim O(g(k))$ if there exists an $M > 0$, such that $|f(k)| \leq M \times g(k)$ for all $k \in \mathbb{N}_0$.

II. Problem Formulation

In this section, we introduce a linear time invariant system model of CPS as well as a replay attack model, which will be employed in the rest of this paper.

Consider the system architecture in Fig. 1. The plant is given by a linear time-invariant system described by the following equation:

$$x_k = Ax_{k-1} + B\phi_k + w_k, \tag{1}$$

where $x_k \in \mathbb{R}^n$ is the state vector at time k, and $w_k \in \mathbb{R}^n$ is a zero mean independently and identically distributed (i.i.d) Gaussian process noise with covariance $Q \geq 0$. $\phi_k \in \mathbb{R}^p$ is the watermark signal that will be discussed in details in Section III.

A sensor network is monitoring the above system. The observation equation is given by

$$y_k = Cx_k + v_k, \tag{2}$$

where $y_k \in \mathbb{R}^m$ is a collection of all sensors' measurements at time k. $v_k \in \mathbb{R}^m$ is a zero mean i.i.d. Gaussian measurement noise with covariance $R \geq 0$. The system parameters A, B and C are unknown.

Remark 1: To simplify notations, in this paper we consider a stable open-loop system. However, our framework can be easily extended to a closed loop system with an unstable plant but a stabilizing controller, which is discussed in Section III.

Notice that the purpose of the watermark signal is intrusion detection instead of stabilization. As a result, we only consider stable systems or systems that have been pre-stabilized by some controller.

We assume that the process noise $w_0, w_1, \cdots$ and the measurement noise $v_0, v_1, \cdots$ are independent of each other. Furthermore, since CPSs usually operate for an extended period of time, it is assumed that the system is already in the steady state, which means that the initial condition x_{-1} is a zero mean Gaussian random vector independent of the process noise and the measurement noise and with covariance Σ, where Σ satisfies the following Lyapunov equation:

$$\Sigma = A\Sigma A^T + Q. \tag{3}$$

We further make the following assumptions regarding the system parameters:

Assumption 1: The system is strictly stable. Furthermore, (A, C) is observable and (A, B) is controllable.

Remark 2: The observability and controllability assumption is without loss of generality as we can perform a Kalman decomposition [21] and only work with the observable and controllable subspace.

Next we introduce a replay attack model. We assume that the adversary has the following capabilities:

1) The attacker has access to all the real-time sensory data. In other words, it knows the sensor's measurements $y_0, \cdots, y_k$.
2) The attacker can modify the real sensor signals y_k to arbitrary sensor signals y'_k.

Given these capabilities, the adversary can employ the following replay attack strategy:

1) The attacker records a sequence of sensor measurements y_ks from time k_1 to $k_1 + T$, where T is large enough to guarantee that the attacker can replay the sequence for an extended period of time during the attack.
2) The attacker modifies the sensor measurements y_k to the recorded signals from time k_2 to $k_2 + T$, i.e., $y'_k = y_{k-\Delta k}$, $\forall\, k_2 \leq k \leq (k_2 + T)$, where $\Delta k = k_2 - k_1$.

Notice that since the system is already in the steady state, both the replayed signal y'_k and the real signal y_k from the sensors will share exactly the same statistics. As a result, replay attack can be stealthy for a large class of linear systems, if no watermark signal is present, i.e. $\phi_k = 0$. For more detailed discussion on the detectability of replay attack, please refer to [13].

The overarching goal of this paper is to design an online learning algorithm for the optimal replay attack detector as well as the optimal covariance U_k of the physical watermark signal, based on the collected input ϕ_k and output y_k, as illustrated in Fig. 1. The physical watermark scheme is introduced in detail in Section III. Based on this scheme, we develop an approach to infer the system parameters based only on the system input data ϕ_k and output data y_k, and design the highlighted parameters in Fig. 1: the covariance U_k of the watermark signal ϕ_k and the optimal detector based on the estimated parameters.

III. Physical Watermark for Systems with Known Parameters

This section introduces the concept of physical watermark, which enables the detection of replay attack. The optimal physical watermark is derived via solving an optimization problem which aims to achieve the optimal trade-off between control performance and intrusion detection. Then we will present the extension to a closed-loop system.

A. Physical Watermark Scheme

The main idea of physical watermark is to inject a random noise ϕ_k, which is called the watermark signal, into the system (1) to excite the system and check whether the system responds to the watermark signal in accordance to the dynamical model of the system. In this section we will restrict the watermark signal ϕ_k to be zero mean i.i.d. Gaussian random variables and its covariance is denoted as U.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2020.2971994, IEEE Transactions on Automatic Control

In the absence of the attack, y_k can be represented as:

$$y_k = \sum_{t=0}^{k} CA^t B\phi_{k-t} + \sum_{t=0}^{k} CA^t w_{k-t} + v_k + CA^{k+1}x_{-1}. \quad (4)$$

For simplicity, let us define

$$\varphi_k \triangleq \sum_{\tau=0}^{k} H_\tau \phi_{k-\tau},\ \vartheta_k \triangleq \sum_{t=0}^{k} CA^t w_{k-t} + v_k + CA^{k+1}x_{-1}, \quad (5)$$

where H_τ is defined as $H_\tau \triangleq CA^\tau B$. Therefore, y_k can be simplified as: $y_k = \varphi_k + \vartheta_k$. It is easy to show that φ_k is a zero mean Gaussian whose covariance converges to $\mathcal{U}$, where $\mathcal{U} \triangleq \sum_{\tau=0}^{\infty} H_\tau U H_\tau^T$. Similarly, ϑ_k is a zero mean Gaussian noise whose covariance is $\mathcal{W} = C\Sigma C^T + R$, where Σ is defined in (3).

On the other hand, let us consider the system under the replay attack, where the replayed y_k' can be written as

$$y_k' = y_{k-\Delta k} = \varphi_{k-\Delta k} + \vartheta_{k-\Delta k}.$$

Now since Δk is unknown to the system operator, we shall treat $\varphi_{k-\Delta k}$ as a zero mean Gaussian random variable with covariance $\mathcal{U}$. As a result, y_k' is a zero mean Gaussian random variable with covariance $\mathcal{U} + \mathcal{W}$. Therefore, to detect replay attack, we need a detector to differentiate the distribution of y_k under the following two hypotheses:

$\mathcal{H}_0$: $y_k \sim \mathcal{N}_0(\varphi_k, \mathcal{W})$, $\mathcal{H}_1$: $y_k \sim \mathcal{N}_1(0, \mathcal{U} + \mathcal{W})$.

Remark 3: It is worth noticing that the watermark signal $\phi_0, \cdots, \phi_k$ are known to the system operator and detector and the conditional distribution (conditioned on $\{\phi_k\}_k$) of y_k converges to a Gaussian distribution with mean φ_k and covariance $\mathcal{W}$.

The NP detector [22] for hypothesis $\mathcal{H}_0$ versus hypothesis $\mathcal{H}_1$ takes the following form:

Lemma 1: At time k, the NP detector rejects $\mathcal{H}_0$ in favor of $\mathcal{H}_1$ if

$$g_k = (y_k - \varphi_k)^T \mathcal{W}^{-1}(y_k - \varphi_k) - y_k^T (\mathcal{W} + \mathcal{U})^{-1} y_k \geq \eta, \quad (6)$$

where η is a threshold chosen by the system operator. Otherwise, hypothesis $\mathcal{H}_0$ is accepted.

Remark 4: For simplicity, we only consider detecting replay attack based on the current measurement y_k. In principle, one may take a moving horizon approach to design a detector, by considering joint distribution of $y_k, y_{k-1}, \cdots, y_{k-\Delta t}$. However, the proposed methodology in this paper can be easily extended to multiple y_ks case by stacking the state vector.

Remark 5: It is worth noticing that since hypothesis $\mathcal{H}_0$ is time-varying due to the φ_k term, the threshold η needs to be time-varying to ensure a constant false alarm rate. If η is still chosen as a constant instead, then the system operator could calculate the expected false alarm rate by numerical integration, since φ_k is a stationary process.

The following theorem quantifies the performance of the detector, in terms of the expected KL-divergence of distribution $\mathcal{N}_0$ and $\mathcal{N}_1$.

Theorem 1: The expected KL divergence of distribution $\mathcal{N}_0$ and $\mathcal{N}_1$ is $\mathbb{E}\, D_{KL}(\mathcal{N}_1 \| \mathcal{N}_0) = \operatorname{tr}\left(\mathcal{U}\mathcal{W}^{-1}\right) - \frac{1}{2}\log\det\left(I + \mathcal{U}\mathcal{W}^{-1}\right)$. Furthermore, the expected KL divergence satisfies the inequality

$$\begin{aligned}\frac{1}{2}\operatorname{tr}\left(\mathcal{U}\mathcal{W}^{-1}\right) &\leq \mathbb{E}\, D_{KL}(\mathcal{N}_1 \| \mathcal{N}_0) \\ &\leq \operatorname{tr}\left(\mathcal{U}\mathcal{W}^{-1}\right) - \frac{1}{2}\log\left[1 + \operatorname{tr}\left(\mathcal{U}\mathcal{W}^{-1}\right)\right].\end{aligned} \quad (7)$$

Proof: The proof is essentially the same as the proof in [14]. ■

Remark 6: It is worth noticing that the expected KL-divergence is a convex function of $\mathcal{U}$ and hence U. However, both the upper and lower bounds of it are increasing functions of $\operatorname{tr}(\mathcal{U}\mathcal{W}^{-1})$. Hence, instead of directly maximizing the detection performance, which is computationally difficult, we could maximize $\operatorname{tr}(\mathcal{U}\mathcal{W}^{-1})$, which is linear with respect to U.

Note that although the watermark signal can enable the detection of replay attack, it also deteriorates the system control performance. As a result, it is important to design the signal to achieve the optimal trade-off between the control performance loss and the detection performance. In this paper, to quantify the performance loss, we use the following Linear Quadratic Gaussian (LQG) metric:

$$J = \lim_{T\to+\infty} \mathbb{E}\left(\frac{1}{T}\sum_{k=0}^{T-1} \begin{bmatrix} y_k \\ \phi_k \end{bmatrix}^T X \begin{bmatrix} y_k \\ \phi_k \end{bmatrix}\right), \quad (8)$$

where $X = \begin{bmatrix} X_{yy} & X_{y\phi} \\ X_{\phi y} & X_{\phi\phi} \end{bmatrix} > 0$ is the weight matrix for the LQG control, which is chosen by the system operator.

Remark 7: The LQG cost is a common choice to quantify the performance of a system running in steady state. On the other hand, we do not foresee any fundamental difficulty to incorporate other performance metrics into our framework, as long as they can be computed from the Markov parameters H_τ.

Since y_k and ϕ_k converge to a stationary process, J can be written in an analytical form as

$$J = \lim_{k\to} \operatorname{tr}\left(X \operatorname{Cov}\left(\begin{bmatrix} y_k \\ \phi_k \end{bmatrix}\right)\right) = \operatorname{tr}\left(X \begin{bmatrix} \mathcal{W} + \mathcal{U} & H_0 U \\ U H_0^T & U \end{bmatrix}\right).$$

Therefore, J is an affine function of U, which can be written as

$$J = J_0 + \Delta J = \operatorname{tr}(X_{yy}\mathcal{W}) + \operatorname{tr}(XS), \quad (9)$$

where J_0 is the optimal LQG cost, and S is linear with respect to U, being defined as $S \triangleq \begin{bmatrix} \mathcal{U} & H_0 U \\ U H_0^T & U \end{bmatrix}$.

Therefore, in order to achieve the optimal trade-off between the control performance and detection performance, we can formulate the following optimization problem:

$$\begin{aligned} U_* = \arg\max_{U \geq 0} &\quad \operatorname{tr}(\mathcal{U}\mathcal{W}^{-1}) \\ \text{subject to} &\quad \operatorname{tr}(XS) \leq \delta, \end{aligned} \quad (10)$$

where δ is a design parameter depending on how much control performance loss is tolerable.

An important property of the optimization problem (10) is that the optimal solution is usually a rank-1 matrix, which is formalized by the following theorem:

Theorem 2: The optimization problem (10) is equivalent to

$$\begin{aligned} U_* = \arg\max_{U \geq 0} &\quad \operatorname{tr}(U\mathcal{P}) \\ \text{subject to} &\quad \operatorname{tr}(U\mathcal{X}) \leq \delta, \end{aligned} \quad (11)$$

where

$$\mathcal{P} \triangleq \sum_{\tau=0}^{\infty} H_\tau^T \mathcal{W}^{-1} H_\tau, \quad (12)$$

$$\mathcal{X} \triangleq \left(\sum_{\tau=0}^{\infty} H_\tau^T X_{yy} H_\tau\right) + H_0^T X_{y\phi} + X_{\phi y} H_0 + X_{\phi\phi}. \quad (13)$$

The optimal solution to (11) is $U_* = zz^T$, where z is the eigenvector corresponding to the maximum eigenvalue of the matrix $\mathcal{X}^{-1}\mathcal{P}$ and $z^T\mathcal{X}z = \delta$. Furthermore, the solution is unique if $\mathcal{X}^{-1}\mathcal{P}$ has only one maximum eigenvalue.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2020.2971994, IEEE Transactions on Automatic Control

Proof: From the definition of $\mathcal{U}$, we know that

$$\begin{aligned}\operatorname{tr}(\mathcal{U}\mathcal{W}^{-1}) &= \sum_{\tau=0}^{\infty} \operatorname{tr}\left(H_\tau U H_\tau^T \mathcal{W}^{-1}\right) \\ &= \sum_{\tau=0}^{\infty} \operatorname{tr}\left(U H_\tau^T \mathcal{W}^{-1} H_\tau\right) = \operatorname{tr}\left(U\mathcal{P}\right).\end{aligned}$$

Following similar steps as in the above proof, we have that $\operatorname{tr}(XS) = \operatorname{tr}(U\mathcal{X})$. Moreover, since $X > 0$, we have that

$$\begin{aligned}\mathcal{X} &\geq H_0^T X_{yy} H_0 + H_0^T X_{y\phi} + X_{\phi y} H_0 + X_{\phi\phi} \\ &\geq X_{\phi\phi} - X_{\phi y} X_{yy}^{-1} X_{y\phi} > 0.\end{aligned}$$

The proof of the second part is similar to the proof of Theorem 7 in [15] and is omitted here due to space limit. ∎

B. Extension to Closed-loop Systems

Before continuing on to the next section, we would like to discuss how to generalize the problem formulation for a closed-loop system with a stabilizing controller. Consider the following system discussed in [13]:

$$x_{k+1} = Ax_k + B(u_k + \phi_k) + w_k,\ y_k = Cx_k + v_k,$$

with the following estimator and controller:

$$\hat{x}_{k+1} = A\hat{x}_k + K(y_{k+1} - CA\hat{x}_k),\ u_k = L\hat{x}_k,$$

and LQG cost as

$$J = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{k=0}^{T-1} y_k^T X_{yy} y_k + (u_k+\phi_k)^T X_{\phi\phi}(u_k+\phi_k)\right],$$

where u_k denotes the optimal LQG control signal.

We can redefine the state $\tilde{x}_k$ and output $\tilde{y}_k$ as $\tilde{x}_k = \begin{bmatrix} x_k \\ \hat{x}_k \end{bmatrix}$, and $\tilde{y}_k = \begin{bmatrix} y_k \\ u_k \end{bmatrix}$, and the design of watermark signal in a closed-loop system can be converted to the open-loop formulation.

It is worth noticing that in order to design the detector and the optimal watermark signal, precise knowledge of the system parameters is needed. However, acquiring the parameters may be troublesome and costly. Furthermore, there may be unforeseen changes in the model of the system, such as topological changes in power systems. As a result, the identified system model may change during the system operation. Therefore, it is beneficial for the system to "learn" the parameters and design the detector and watermark signal in real-time, which will be our focus in the next section.

IV. Physical Watermark for Systems with Unknown Parameters

This section is devoted to developing an online "learning" procedure to infer the system parameters, based on which, we show how to design watermark signals and the optimal detector and prove that the physical watermark and the detector asymptotically converge to the optimal ones.

Throughout the section, we make the following assumptions:

Assumption 2: 1) A is diagonalizable. 2) The maximum eigenvalue of $\mathcal{X}^{-1}\mathcal{P}$ is unique. 3) The system is not under attack during the learning phase. 4) The number of distinct eigenvalues of A, which is denoted as $\tilde{n}$, is known. 5) The LQG weight matrix X and the largest tolerable LQG loss δ are known.

Remark 8: The first and second assumptions are required in order to ensure that the optimal covariance of the watermark signal is a differentiable function of H_τ, i.e., the problem is not ill-conditioned. The third assumption is necessary since there is no way to do system identification without (real) sensory data and it is also needed to prove the asymptotic convergence of our algorithm to the true optimal solution as this cannot be achieved in finite time due to the inherent process and measurement noises. Nevertheless, we shall illustrate through simulation, that after a certain period of learning phase, our algorithm can approximate the optimal solution with reasonably good accuracy and the system can detect replay attack. The fourth assumption is also required to prove convergence, although we shall demonstrate in the simulation that we can use a reduced model to approximate the system with good accuracy. The fifth assumption should hold for all practical cases as X and δ are design parameters chosen by the system operator.

For the sake of legibility, we shall introduce our algorithm first and present the theorem on the correctness of our approach in the end.

A. An Online Algorithm

In this subsection, we will present the complete algorithm in a pseudo-code form. After that, the online "learning" scheme will be introduced in detail.

Algorithm 1 describes our proposed online watermarking algorithm. The notations are described later in the subsection.

First, we initialize some parameters which will be used later. In each round of the **while** iteration, the optimal covariance of the watermarking $U_{k,*}$ based on current knowledge is computed firstly. Based on the derived covariance, one can update the covariance U_k by combining "exploration" and "exploitation" term which will be described in detail later. According to the updated covariance, we generate the watermarking signals ϕ_k and inject them to the plant. Then we collect the sensory data y_k and employ them and watermarking signals to infer necessary system parameters $H_{k,\tau}, \mathcal{P}_k, \mathcal{X}_k$. Based on the estimated parameters, one can update the NP detector $\hat{g}_k$. Then one can repeat the above process to identify system parameters and design the watermarking signals as well as the optimal detector. A pseudo-code form for Algorithm 1 is as follows:

Algorithm 1 Online Watermarking Design

```
Initialization: $\mathcal{P}_{-1} \leftarrow I,\ \mathcal{X}_{-1} \leftarrow X_{\phi\phi}, k \leftarrow 0$
Iteration:
 1: while true do
 2:     $U_{k,*} \leftarrow \arg\max_{U\geq 0,\, \operatorname{tr}(U\mathcal{X}_{k-1})\leq\delta} \operatorname{tr}(U\mathcal{P}_{k-1})$
 3:     $U_k \leftarrow U_{k,*} + (k+1)^{-\beta}\delta I$
 4:     Generate random variable $\zeta_k \sim \mathcal{N}(0, I)$
 5:     Apply watermark signal $\phi_k \leftarrow U_k^{1/2}\zeta_k$
 6:     Collect sensory data $y_k$
 7:     $H_{k,\tau} \leftarrow \frac{1}{k-\tau+1}\sum_{t=\tau}^{k} y_t\phi_{t-\tau}^T U_{t-\tau}^{-1}$
 8:     Compute the coefficient of $p_k(x)$ by solving (18)
 9:     if $p_k(x)$ is Schur stable then
10:         Update $\mathcal{P}_k, \mathcal{X}_k$ from (19)-(21)
11:     end if
12:     Update $\hat{g}_k$ from (22)
13:     $k \leftarrow k+1$
14: end while
```

Remark 9: For Algorithm 1, $\mathcal{P}_k, \mathcal{X}_k$ are defined in (15), U_k is the covariance of watermarking signal, and $H_{k,\tau}$ is defined in (16). Step 3 is the update of the covariance of the physical watermark in (14). All parameters will be illustrated in the following subsections.

Then we will introduce this algorithm in detail.

Generation of the Watermark Signal ϕ_k: Let us design U_k, which can be considered as an approximation for the optimal co-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2020.2971994, IEEE Transactions on Automatic Control

variance of the watermark signal U, as

$$U_k = U_{k,*} + \frac{\delta}{(k+1)^{\beta}} I, \tag{14}$$

where $0 < \beta < 1$, δ is the maximum tolerable LQG loss defined in (10), and $U_{k,*}$ is the solution of the following optimization problem

$$\begin{aligned} U_{k,*} = &\arg\max_{U \geq 0} && \operatorname{tr}(U\mathcal{P}_{k-1}), \\ &\text{subject to} && \operatorname{tr}(U\mathcal{X}_{k-1}) \leq \delta, \end{aligned} \tag{15}$$

and $\mathcal{P}_{k-1}$ and $\mathcal{X}_{k-1}$ are the estimates of $\mathcal{P}$ and $\mathcal{X}$ matrices, respectively, based on $y_0, \dots, y_{k-1}, \phi_0, \dots, \phi_{k-1}$, both of which are initialized as: $\mathcal{P}_{-1} = I$, $\mathcal{X}_{-1} = X_{\phi\phi}$. The inference procedure of $\mathcal{P}_k$ and $\mathcal{X}_k$ for $k \geq 0$ will be provided in the further subsections.

Remark 10: Notice that the second term $(k+1)^{-\beta} I$ on the RHS of (14) is crucial for parameter identification. The reason is that $U_{k,*}$ is in general a rank 1 matrix (as is proved in Thereom 2) and hence it does not provide persistent excitation to the system for us to identify the necessary parameters. Conceptually, the $(k+1)^{-\beta} I$ term can be interpreted as an "exploration" term, as it provides necessary excitation to the system in order for us to infer the parameters. The $U_{k,*}$ is the "exploitation" term, as it is optimal under our current knowledge of the system parameters.

At each time k, the watermark signal is chosen to be $\phi_k = U_k^{1/2} \zeta_k$, where ζ_ks are i.i.d. Gaussian random vectors with covariance I.

The rest of this section is devoted to inferencing the system parameters from the collected sensory data $y_0, \dots, y_k$ and watermarks $\phi_0, \dots, \phi_k$. We will first identify the Markov parameters H_τ.

Inference on H_τ: Let us define $H_{k,\tau}$, where $0 \leq \tau \leq 3\tilde{n} - 2$, as

$$\begin{aligned} H_{k,\tau} &\triangleq \frac{1}{k-\tau+1} \sum_{t=\tau}^{k} y_t \phi_{t-\tau}^T U_{t-\tau}^{-1} \\ &= H_{k-1,\tau} + \frac{1}{k-\tau+1} \left(y_k \phi_{k-\tau}^T U_{k-\tau}^{-1} - H_{k-1,\tau} \right), \end{aligned} \tag{16}$$

where $H_{k,\tau}$ is an estimate of H_τ at time k.

Remark 11: It is worth noticing that other methods, such as subspace identification, may be superior for classical system identification tasks to the method we proposed. However, since the covariance of our watermark signal converges to a degenerate matrix (of rank 1), it is non-trivial to analyze the convergence properties for more advanced system identification methods, such as subspace identification, which we shall leave as a further research direction.

It is worth noticing that the calculation of the matrices $\mathcal{U}$, $\mathcal{W}$, $\mathcal{P}$ and $\mathcal{X}$ requires H_τ for all $\tau \geq 0$. Next we shall show that in fact only finitely many H_τs are needed to compute those matrices, which requires one intermediate result:

Lemma 2: Assuming the matrix A is diagonalizable with $\lambda_1, \dots, \lambda_{\tilde{n}}$ being its distinct eigenvalues, then there exist unique $\Omega_1, \cdots, \Omega_{\tilde{n}}$, such that $H_\tau = \sum_{i=1}^{\tilde{n}} \lambda_i^\tau \Omega_i$.

Proof: Without loss of generality, we assume that A is a diagonal matrix. As a result, $A^\tau = \operatorname{diag}(\lambda_1^\tau I_1, \cdots, \lambda_{\tilde{n}}^\tau I_{\tilde{n}})$, where λ_i denotes the ith distinct eigenvalue of A, λ_i^τ denotes λ_i to the power of τ, and I_i is the identity matrix of size n_i by n_i with n_i the multiplicity of λ_i. Hence, we have $H_\tau = \sum_{i=1}^{\tilde{n}} \lambda_i^\tau \Omega_i$, with $\Omega_i = C\operatorname{diag}(0, \dots, 0, I_i, 0, \dots, 0)B$, which completes the proof. ■

Since A satisfies its own minimal polynomial $p(x) = \prod_{i=1}^{\tilde{n}}(x - \lambda_i) = x^{\tilde{n}} + \alpha_{\tilde{n}-1} x^{\tilde{n}-1} + \ldots + \alpha_0$, we know that for any $i \geq 0$:

$$H_{i+\tilde{n}} + \alpha_{\tilde{n}-1} H_{i+\tilde{n}-1} + \cdots + \alpha_0 H_i = CA^i p(A) B = 0. \tag{17}$$

Leveraging (17), we could use $H_0, H_1, \cdots, H_{3\tilde{n}-2}$ to estimate both λ_is and Ω_is and thus H_τ for any τ. To this end, let us define:

$$\begin{bmatrix} \alpha_{k,0} \\ \vdots \\ \alpha_{k,\tilde{n}-1} \end{bmatrix} \triangleq -\Xi_k^{-1} \begin{bmatrix} \operatorname{tr}(\mathcal{H}_{k,0}^T \mathcal{H}_{k,\tilde{n}}) \\ \vdots \\ \operatorname{tr}(\mathcal{H}_{k,\tilde{n}-1}^T \mathcal{H}_{k,\tilde{n}}) \end{bmatrix}, \tag{18}$$

where $\Xi_k \triangleq \begin{bmatrix} \operatorname{tr}(\mathcal{H}_{k,0}^T \mathcal{H}_{k,0}) & \cdots & \operatorname{tr}(\mathcal{H}_{k,0}^T \mathcal{H}_{k,\tilde{n}-1}) \\ \vdots & \ddots & \vdots \\ \operatorname{tr}(\mathcal{H}_{k,\tilde{n}-1}^T \mathcal{H}_{k,0}) & \cdots & \operatorname{tr}(\mathcal{H}_{k,\tilde{n}-1}^T \mathcal{H}_{k,\tilde{n}-1}) \end{bmatrix}$, and $\mathcal{H}_{k,i} \triangleq \begin{bmatrix} H_{k,i} \\ H_{k,i+1} \\ \vdots \\ H_{k,i+2\tilde{n}-2} \end{bmatrix}$.

Remark 12: One can prove that $\alpha_{k,i}$ from (18) is the solution of the minimization problem: $\min \|\mathcal{H}_{k,\tilde{n}} + \alpha_{\tilde{n}-1}\mathcal{H}_{k,\tilde{n}-1} + \cdots + \alpha_0 \mathcal{H}_{k,0}\|_F$, where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix.

Let us denote the roots of the polynomial $p_k(x) = x^{\tilde{n}} + \alpha_{k,\tilde{n}-1} x^{\tilde{n}-1} + \cdots + \alpha_{k,0}$ to be $\lambda_{k,1}, \cdots, \lambda_{k,\tilde{n}}$. Define a Vandermonde like matrix V_k to be

$$V_k \triangleq \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \lambda_{k,1} & \lambda_{k,2} & \cdots & \lambda_{k,\tilde{n}} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{k,1}^{3\tilde{n}-2} & \lambda_{k,2}^{3\tilde{n}-2} & \cdots & \lambda_{k,\tilde{n}}^{3\tilde{n}-2} \end{bmatrix},$$

where $\lambda_{k,i}$ is an estimate of λ_i at time k and $\lambda_{k,i}^\tau$ is $\lambda_{k,i}$ to the power of τ, and we shall estimate Ω_i as

$$\begin{bmatrix} \Omega_{k,1} \\ \vdots \\ \Omega_{k,\tilde{n}} \end{bmatrix} = (V_k \otimes I_m)^+ \begin{bmatrix} H_{k,0} \\ \cdots \\ H_{k,3\tilde{n}-2} \end{bmatrix}. \tag{19}$$

Inference on φ_k, ϑ_k and $\mathcal{W}$: This subsection is devoted to the inference of φ_k and ϑ_k defined in (5), which corresponds to the parts of y_k generated by the watermark signal and noise respectively. We will further infer the covariance $\mathcal{W}$ of ϑ_k.

Let us define $\hat{\varphi}_k \triangleq \sum_{i=1}^{\tilde{n}} \hat{\varphi}_{k,i}$, with $\hat{\varphi}_{k,i} = \lambda_{k,i}\hat{\varphi}_{k-1,i} + \Omega_{k,i}\phi_k$, and $\hat{\varphi}_{-1,i} = 0$. As a result, we can estimate ϑ_k as $\hat{\vartheta}_k \triangleq y_k - \hat{\varphi}_k$. The covariance of ϑ_k can be estimated as $\mathcal{W}_k \triangleq \frac{1}{k+1} \sum_{t=0}^{k} \hat{\vartheta}_t \hat{\vartheta}_t^T$.

Inference on $\mathcal{P}$, $\mathcal{X}$, $\mathcal{U}$ and g_k: Finally we can derive an estimation of the $\mathcal{P}$ and $\mathcal{X}$ matrices, which are required to compute the optimal covariance U of the watermark signal, given by

$$\begin{aligned} \mathcal{P}_k &= \sum_{\tau=0}^{\infty} \left(\sum_{i=1}^{\tilde{n}} \lambda_{k,i}^\tau \Omega_{k,i} \right)^T \mathcal{W}_k^{-1} \left(\sum_{i=1}^{\tilde{n}} \lambda_{k,i}^\tau \Omega_{k,i} \right) \\ &= \sum_{\tau=0}^{\infty} \left(\sum_{i=1}^{\tilde{n}} \sum_{j=1}^{\tilde{n}} \lambda_{k,i}^\tau \lambda_{k,j}^\tau \Omega_{k,i}^T \mathcal{W}_k^{-1} \Omega_{k,j} \right) \\ &= \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{\tilde{n}} \left(\sum_{\tau=0}^{\infty} \left(\lambda_{k,i} \lambda_{k,j} \right)^\tau \right) \Omega_{k,i}^T \mathcal{W}_k^{-1} \Omega_{k,j} \\ &= \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{\tilde{n}} \frac{1}{1 - \lambda_{k,i}\lambda_{k,j}} \Omega_{k,i}^T \mathcal{W}_k^{-1} \Omega_{k,j}, \end{aligned} \tag{20}$$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2020.2971994, IEEE Transactions on Automatic Control

where (20) is derived from the summation of geometric series, and

$$\begin{aligned}\mathcal{X}_k =& \sum_{\tau=0}^{\infty}\left(\sum_{i=1}^{\tilde n}\lambda_{k,i}^{\tau}\Omega_{k,i}\right)^T X_{yy}\left(\sum_{i=1}^{\tilde n}\lambda_{k,i}^{\tau}\Omega_{k,i}\right)\\ &+\sum_{i=1}^{\tilde n}\Omega_{k,i}^T X_{y\phi}+X_{\phi y}\sum_{i=1}^{\tilde n}\Omega_{k,i}+X_{\phi\phi}\\ =&\sum_{i=1}^{\tilde n}\sum_{j=1}^{\tilde n}\frac{1}{1-\lambda_{k,i}\lambda_{k,j}}\Omega_{k,i}^T X_{yy}\Omega_{k,j}+\sum_{i=1}^{\tilde n}\Omega_{k,i}^T X_{y\phi}\\ &+X_{\phi y}\sum_{i=1}^{\tilde n}\Omega_{k,i}+X_{\phi\phi}. \end{aligned} \tag{21}$$

The NP detection statistics g_k can be approximated by

$$\hat g_k = (y_k-\hat\varphi_k)^T\mathcal{W}_k^{-1}(y_k-\hat\varphi_k)-y_k^T(\mathcal{W}_k+\mathcal{U}_k)^{-1}y_k, \tag{22}$$

where

$$\begin{aligned}\mathcal{U}_k =& \sum_{\tau=0}^{\infty}\left(\sum_{i=1}^{\tilde n}\lambda_{k,i}^{\tau}\Omega_{k,i}\right)U_{k,*}\left(\sum_{i=1}^{\tilde n}\lambda_{k,i}^{\tau}\Omega_{k,i}\right)^T\\ =&\sum_{i=1}^{\tilde n}\sum_{j=1}^{\tilde n}\frac{1}{1-\lambda_{k,i}\lambda_{k,j}}\Omega_{k,i}U_{k,*}\Omega_{k,j}^T. \end{aligned} \tag{23}$$

Remark 13: For the proposed online algorithm, the system identification and watermark design are tightly coupled. As is commented in Remark 10, the watermarking-based replay attack detection requires the injection of a rank-1 watermarking signal (assuming it is performed optimally). On the other hand, persistency of excitation is required for system identification, i.e., the injected signal needs to be full rank. As a result, we carefully design the covariance of the injected signal to be the "optimal" rank-1 covariance matrix on the current knowledge of the system, plus a diminishing factor $(k+1)^{-\beta}I$, and we further prove that this additional term, although vanishing asymptotically, provides us with enough information to perfectly identify the necessary parameters of the system.

Remark 14: It is worth noticing that comparing to an approach with off-line system identification first and then watermarking design later, our approach provides three advantages: 1) Finite-time system identification cannot identify the system parameters precisely and hence the watermarking scheme will not be optimal if the system identification process is stopped. 2) In practice, the control system could slowly change due to various reasons (e.g., components wear out), so we need to adjust the parameters continuously. 3) For many practical control systems, a model of the system is not available. It is often expensive to stop the system operation to perform off-line system identification.

B. Algorithm Properties

The following theorem establishes the convergence of $U_{k,*}$ and $\hat g_k$, the proof of which is reported in the appendix.

Theorem 3: Assuming that A is strictly stable and Assumption 2 holds. If $0<\beta<1$, then for any $\epsilon>0$, the following limits hold almost surely:

$$\lim_{k\to\infty}\frac{U_{k,*}-U_*}{k^{-\gamma+\epsilon}}=0,\ \lim_{k\to\infty}\frac{\hat g_k-g_k}{k^{-\gamma+\epsilon}}=0, \tag{24}$$

where $\gamma=(1-\beta)/2>0$. In particular, $U_{k,*}$ and $\hat g_k$ almost surely converge to U_* and g_k respectively.

From the definition of $U_k=U_{k,*}+(k+1)^{-\beta}\delta I$, we immediately have the following corollary:

Corollary 1: Assume that A is strictly stable and Assumption 2 holds. If $0<\beta<1$, then for any $\epsilon>0$, $\lim_{k\to\infty}\frac{U_k-U_*}{k^{-\min(\gamma,\beta)+\epsilon}}=0$ holds almost surely.

Remark 15: It is worth noticing that (24) implies that both $U_{k,*}-U_*$ and $\hat g_k-g_k$ are of the order $O(k^{-\gamma+\epsilon})$ as k goes to infinity. Hence, the convergence rate γ is maximized when $\beta\to 0^+$, which corresponds to the case where the exploration term $(k+1)^{-\beta}\delta I$ in U_k stays constant. However, although this will maximize the performance for the inference algorithm, the covariance U_k of the watermark signal ϕ_k will not converge to the true optimal U_*. In order to achieve "fastest" convergence rate of U_k, we need to choose the decay rate for the exploration term to be $\beta=1/3=\arg\max_\beta(\gamma,\beta)$.

We would also like to point out that Theorem 3 only provides an upper bound for the almost sure convergence rate and we plan to investigate the exact convergence rate in our future work. It is also interesting to see if faster convergence can be achieved by using more advanced system identification techniques.

V. Simulation Result

In this section, the performance of the proposed algorithm is evaluated. We will apply the proposed online "learning" approach to an industrial process, Tennessee Eastman Process (TEP).

Tennessee Eastman Process (TEP) is a commonly used process control system proposed by Downs and Vogel in [23]. In this simulation, we adopt a simplified version of TEP from [24]: $\dot x = Ax+Bu$, $y=Cx$, where A, B and C are constant matrices [1].

This system simulates a MIMO system of order $n=8$ with $p=4$ inputs and $m=10$ outputs. We discretize the system using the control system toolbox in MATLAB, by selecting a sample time of $0.6s$. Again, we choose X in (8), the covariance matrices Q and R to be identity matrices with proper dimensions. We assume that δ in (11) is equal to 5% of J_0, and $\beta=1/3$. In this simulation, we assume that we do not know the dimension of the state space, which is 8, and instead we underestimate it by assuming that A only has $\tilde n=5$ distinct eigenvalues.

Fig. 2 illustrates the relative error $\|U_{k,*}-U_*\|_F/\|U_*\|_F$ after running the system for roughly 1 week ($10^6\times 0.6s\approx 0.992$week). Fig. 3 illustrates the NP statistics g_k and the estimated NP statistics $\hat g_k$, assuming that the adversary collects the measurement from 10^6+1 to 10^6+100 and replays them to the system from time 10^6+101 to 10^6+200. One can see that although we underestimate the dimensions of the system, our algorithm can still achieve a high accuracy.

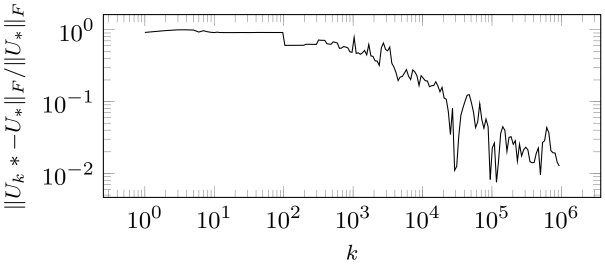

Fig. 2. Relative error of $\boldsymbol{U_{k,*}}$.

VI. Conclusion

In this paper, an algorithm that can simultaneously generate the watermarking signal and infer the system parameters required for

[1] For more details about this dynamic model, please refer to Appendix I in [24].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2020.2971994, IEEE Transactions on Automatic Control

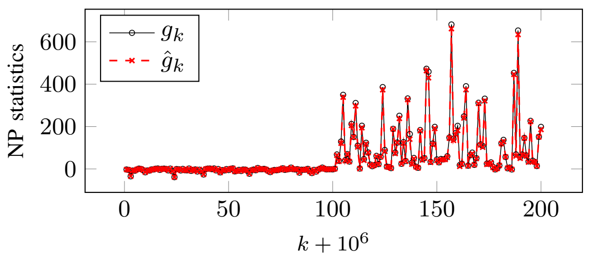

Fig. 3. The detection statistics v.s. time. The black solid line with circle markers is the true NP statistics $\boldsymbol{g_k}$, assuming full system knowledge. The red dashed line with cross markers denotes our estimated $\hat{\boldsymbol{g}}_{\boldsymbol{k}}$.

optimal replay attack detection is proposed. We prove that our algorithm converges to the optimal one and characterize an upper bound for the almost surely convergence rate. For future works, we would like to quantify the exact convergence rate, as well as exploring other system identification methods and prove their convergence. It is of interest to study secure control in other cases, such as batach-operating process. We are also interested in adversarial learning when the sensor data is compromised.

Appendix

The whole appendix is devoted to proving Theorem 3. We shall present several preliminary results first and then proceed with the proof of Theorem 3. Please refer to [27] for more details of the proof.

Preliminary Results: To simplify notations, for a random variable (vector, or matrices) x_k, we denote that $x_k \sim \mathcal{C}(\alpha)$ if for all $\epsilon > 0$, we have that $x_k \sim O(k^{\alpha+\epsilon})$, i.e., $\lim_{k\to\infty} \frac{\|x_k\|}{k^{\alpha+\epsilon}} \overset{a.s.}{=} 0$.

Notice that $x_k \sim O(k^\alpha)$ implies that $x_k \sim \mathcal{C}(\alpha)$, but the reverse is not necessarily true[2]. The following lemma establishes some basic properties of $\mathcal{C}(\alpha)$ functions:

Lemma 3: Assuming that $x_k \sim \mathcal{C}(\alpha)$ and $y_k \sim \mathcal{C}(\beta)$, with $\alpha \geq \beta$, then: 1) $x_k + y_k \sim \mathcal{C}(\alpha)$, $x_k \times y_k \sim \mathcal{C}(\alpha+\beta)$, and $(x_k + \Delta x_k)(y_k + \Delta y_k) - x_k y_k \sim \mathcal{C}(\max\{\alpha\beta', \alpha'\beta, \alpha'\beta'\})$, suppose that $\Delta x_k \sim \mathcal{C}(\alpha')$ and $\Delta y_k \sim \mathcal{C}(\beta')$. 2) $\sum_{t=0}^k x_t \sim \mathcal{C}(\alpha+1)$. 3) Suppose f is differentiable at 0 and $\alpha < 0$, then $f(x_k) - f(0) \sim \mathcal{C}(\alpha)$. 4) $s_k \sim \mathcal{C}(\alpha)$, with $s_k = \rho s_{k-1} + x_k$, $s_{-1} = 0$, where $|\rho| < 1$. 5) Assume that X_k is a matrix and $X_k \sim \mathcal{C}(\alpha)$. Let $S_k = AS_{k-1}B + X_k$, $S_{-1} = 0$, where A, B are matrices of proper dimensions. Then $S_k \sim \mathcal{C}(\alpha)$ if $B^T \otimes A$ is strictly stable. 6) $\zeta_k \sim \mathcal{C}(0)$, where $\{\zeta_k\}$ is a sequence of i.i.d. Gaussian random variables, i.e., $\zeta_k \sim \mathcal{N}(\bar{\mu}, Z)$.

Let $\{\mathcal{F}_k\}$ be a filtration of sigma algebras and $\{M_k\}$ be a matrix-valued stochastic process that is adapted to the filtration $\{\mathcal{F}_k\}$, we call $\{M_k\}$ a (matrix-valued) martingale (with respect to the filtration $\{\mathcal{F}_k\}$) if for all t, $\mathbb{E}\left(M_{k+1}|\mathcal{F}_k\right) = M_k$ holds.

For the rest of the paper, we shall assume that the filtration $\mathcal{F}_k$ is the σ-algebra generated by the random variables $\{x_{-1}, \phi_0, \cdots, \phi_k, w_0, \cdots, w_k, v_0, \cdots, v_k\}$. Now we have the following lemma to establish a strong law for matrix-valued martingale:

Lemma 4: If $M_k = \Phi_0 + \Phi_1 + \cdots + \Phi_k$ is a matrix-valued martingale such that $\mathbb{E}\,\|\Phi_k\|^2 \sim \mathcal{C}(\beta)$, where $0 \leq \beta < 1$, then M_k/k converges to 0 almost surely. Furthermore, $\frac{M_k}{k} \sim \mathcal{C}\left(\frac{\beta-1}{2}\right)$.

Lemma 5: U_k is upper and lower bounded by $\delta(k+1)^{-\beta} I \leq U_k \leq \delta\left(\left(X_{\phi\phi} - X_{\phi y}X_{yy}^{-1}X_{y\phi}\right)^{-1} + I\right)$.

We are now ready to prove Theorem 3, which requires several intermediate steps.

Lemma 6: $H_{k,\tau} - H_\tau \sim \mathcal{C}(-\gamma)$, with $\gamma = (1-\beta)/2$. In particular, $H_{k,\tau}$ converges to H_τ almost surely.

Proof: It is easy to see that y_k and U_{k+1} are measurable w.r.t. $\mathcal{F}_k$. Furthermore, let $k_1, k_2 \geq 0$ be two time indices, then one have

$$\mathbb{E}(\phi_{k_1}\phi_{k_2+1}^T|\mathcal{F}_{k_2}) = \begin{cases} U_{k_2+1} & \text{if } k_1 = k_2+1 \\ 0 & \text{otherwise} \end{cases},$$
$$\mathbb{E}(w_{k_1}\phi_{k_2+1}^T|\mathcal{F}_{k_2}) = 0,\ \mathbb{E}(v_{k_1}\phi_{k_2+1}^T|\mathcal{F}_{k_2}) = 0, \tag{25}$$

which, combined with (4), implies that

$$\mathbb{E}\left(y_{k+\tau}\phi_k^T U_k^{-1}|\mathcal{F}_{k-1}\right) = H_\tau. \tag{26}$$

Next we shall compute the expectation of $\|y_{k+\tau}\phi_k^T U_k^{-1}\|^2$. Notice that $\phi_k = U_k^{1/2}\zeta_k$, where ζ_k follows the standard normal distribution. Hence,

$$\begin{aligned}\|y_{k+\tau}\phi_k^T U_k^{-1}\|^2 &= \|y_{k+\tau}\phi_k^T U_k^{-2}\phi_k y_{k+\tau}^T\| \\ &\leq \|y_{k+\tau}\|^2\|\zeta_k\|^2\|U_k^{-1}\| \leq \delta(k+1)^\beta\|y_{k+\tau}\|^2\|\zeta_k\|^2.\end{aligned}$$

The last inequality is true due to Lemma 5. As a result, by Cauchy-Schwarz inequality, $\mathbb{E}\|y_{k+\tau}\phi_k^T U_k^{-1}\|^2 \leq \delta(k+1)^\beta\sqrt{\mathbb{E}\|y_{k+\tau}\|^4}\sqrt{\mathbb{E}\|\zeta_k\|^4}$.

Notice that $\|\zeta_k\|$ is χ-distributed with p degree of freedom, which implies that $\mathbb{E}\|\zeta_k\|^4 = p(p+2)$. On the other hand, one can prove that $\sup_k \mathbb{E}\|y_k\|^4$ is bounded since by 5, U_k is upper bounded. As a result, we prove that $\mathbb{E}\|y_{k+\tau}\phi_k^T U_k^{-1}\|^2 \sim \mathcal{C}(\beta)$, which further implies that

$$\begin{aligned}\mathbb{E}\|y_{k+\tau}\phi_k^T U_k^{-1} - H_\tau\|^2 &\leq \mathbb{E}\left(\|y_{k+\tau}\phi_k^T U_k^{-1}\| + \|H_\tau\|\right)^2 \\ &\leq \mathbb{E}\left(2\|y_{k+\tau}\phi_k^T U_k^{-1}\|^2 + 2\|H_\tau\|^2\right) \sim \mathcal{C}(\beta).\end{aligned} \tag{27}$$

As a result, by (26), one can prove that the following stochastic process is a matrix-valued martingale

$$\mathcal{S}_{\tau,i}(k+1) = \mathcal{S}_{\tau,i}(k) + \left[y_{(k+1)\tilde{\tau}+i}\phi_{k\tilde{\tau}+i+1}U_{k\tilde{\tau}+i+1}^{-1} - H_\tau\right]$$

for the filtration $\mathcal{F}_{k\tilde{\tau}+i}$, where $\tilde{\tau} = \tau + 1$, and $0 \leq i \leq \tau$. Now by (27) and Lemma 4, we know that $\frac{\mathcal{S}_{\tau,i}(k)}{k} \sim \mathcal{C}(-\gamma)$. From the definition of $\mathcal{S}_{\tau,i}(k)$, one can see that for large enough k,

$$H_{k,\tau} - H_\tau = \sum_{i=0}^{\tau}\frac{k_i}{k} \times \frac{\mathcal{S}_{\tau,i}(k_i)}{k_i}, \tag{28}$$

where $k_i = \max\{t \in \mathbb{N} : t\tilde{\tau} + i \leq k\}$. Notice that $k_i \geq 0$ and $\sum k_i = k$. Hence, the estimation error of $H_{k,\tau} - H_\tau$ is a convex combination of $\mathcal{S}_{\tau,i}$s. As a result, for any $\epsilon > 0$,

$$\|H_{k,\tau} - H_\tau\| \leq \max_{0\leq i\leq\tau}\frac{\|\mathcal{S}_{\tau,i}(k_i)\|}{k_i} \sim O(k_i^{-\gamma+\epsilon}). \tag{29}$$

Notice that when k is large enough, $k/k_i \to \tau$, which implies that $H_{k,\tau} - H_\tau \sim \mathcal{C}(-\gamma)$. The a.s. convergence can be trivially proved by the fact that $\gamma > 0$ is positive. ■

By Lemma 2, 3.3, 6 and the fact that (A, B) is controllable and (A, C) is observable, one can derive that $\lambda_{k,i} - \lambda_i \sim \mathcal{C}(-\gamma)$ and $\Omega_{k,i} - \Omega_i \sim \mathcal{C}(-\gamma)$.

Notice that the error between $\hat{\varphi}_{k,i}$ and $\varphi_{k,i}$ satisfies the recursive equation: $\varphi_{k+1,i} - \hat{\varphi}_{k+1,i} = (\lambda_i - \lambda_{k,i})\varphi_{k,i} + \lambda_{k,i}(\varphi_{k,i} - \hat{\varphi}_{k,i}) + (\Omega_i - \Omega_{k,i})\phi_k$. For any $\epsilon > 0$, we have

$$\begin{aligned}\frac{\|\varphi_{k+1,i} - \hat{\varphi}_{k+1,i}\|}{(k+1)^{-\gamma+2\epsilon}} \leq |\lambda_{k,i}|\frac{\|\varphi_{k,i} - \hat{\varphi}_{k,i}\|}{k^{-\gamma+2\epsilon}} + \frac{|\lambda_i - \lambda_{k,i}|}{k^{-\gamma+\epsilon}}\frac{\|\varphi_{k,i}\|}{k^\epsilon} \\ + \frac{\|\Omega_i - \Omega_{k,i}\|}{k^{-\gamma+\epsilon}}\frac{\|\phi_k\|}{k^\epsilon}.\end{aligned}$$

[2] To see a counterexample, $\log k \sim \mathcal{C}(0)$, but $\log k$ is not of the order $O(k^0)$.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2020.2971994, IEEE Transactions on Automatic Control

Notice that $\phi_k = U_k^{1/2}\zeta_k$. Since $\zeta_k \sim \mathcal{C}(0)$ by Lemma 3.6, and U_k is upper bounded by Lemma 5, $\phi_k \sim \mathcal{C}(0)$. Thus, $\varphi_{k,i} \sim \mathcal{C}(0)$ by Lemma 3.4. Furthermore, since $\lambda_{k,i} - \lambda_i \sim \mathcal{C}(-\gamma)$ and $\Omega_{k,i} - \Omega_i \sim \mathcal{C}(-\gamma)$, for any $\epsilon_1 > 0$, there exists K (possibly random), such that for any $k \geq K$, the following inequalities hold almost surely,

$$|\lambda_i - \lambda_{k,i}| \leq \epsilon_1, \; \frac{|\lambda_i - \lambda_{k,i}|}{k^{-\gamma+\epsilon}} \frac{\|\varphi_{k,i}\|}{k^{\epsilon}} + \frac{\|\Omega_i - \Omega_{k,i}\|}{k^{-\gamma+\epsilon}} \frac{\|\phi_k\|}{k^{\epsilon}} \leq \epsilon_1.$$

Therefore, for $k \geq K$, we have

$$\frac{\|\varphi_{k+1,i} - \hat{\varphi}_{k+1,i}\|}{(k+1)^{-\gamma+2\epsilon}} \leq (|\rho| + \epsilon_1) \times \frac{\|\varphi_{k,i} - \hat{\varphi}_{k,i}\|}{k^{-\gamma+2\epsilon}} + \epsilon_1. \; a.s.$$

Now since $|\rho| < 1$, we can choose ϵ_1 small enough such that $|\rho| + \epsilon_1 < 1$, therefore, $\limsup_{k\to\infty} \frac{\|\varphi_{k,i} - \hat{\varphi}_{k,i}\|}{k^{-\gamma+2\epsilon}} \leq \frac{\epsilon_1}{1-|\rho|-\epsilon_1}. \; a.s.$ Hence, $\|\varphi_{k,i} - \hat{\varphi}_{k,i}\|/k^{-\gamma+3\epsilon} \overset{a.s.}{\to} 0$, which proves that $\varphi_{k,i} - \hat{\varphi}_{k,i} \sim \mathcal{C}(-\gamma)$.

Lemma 7: $\frac{1}{k+1}\sum_{t=0}^{k} \vartheta_t \vartheta_t^T - \mathcal{W} \sim \mathcal{C}(-0.5)$, where $\vartheta_k \triangleq \sum_{t=0}^{k} CA^t w_{k-t} + v_k + CA^{k+1}x_{-1}$.

Proof: Let us define function $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, such that for any symmetric matrix $X \in \mathbb{S}^{n\times n}$,

$$\mathcal{A}(X) = X + AXA^T + A^2XA^{2T} + \cdots ,$$

For non-symmetric $X \in \mathbb{R}^{n\times n}$, we define $\mathcal{A}(X) = \mathcal{A}\left(\frac{X+X^T}{2}\right)$. One can prove that $\mathcal{A}(X) - A^k\mathcal{A}(X)A^{kT} = \sum_{i=0}^{k-1} A^i \frac{X+X^T}{2} A^{iT}$. To simplify notations, let us define $w_{-1} = x_{-1}$. By mathematical induction, $\sum_{t=0}^{k} \vartheta_t\vartheta_t^T = \mathcal{M}_k - CA\mathcal{N}_kA^TC^T$, where

$$\mathcal{M}_k = \mathcal{M}_{k-1} + \Pi_k$$

$$\mathcal{N}_k = A\mathcal{N}_{k-1}A^T + 2\mathcal{A}\left(\left(\sum_{t=-1}^{k} A^{k-t}w_t\right) w_k^T\right) - \mathcal{A}(w_kw_k^T).$$

with

$$\Pi_k = v_kv_k^T + v_k\left(\sum_{t=-1}^{k} CA^{k-t}w_t\right)^T + \left(\sum_{t=-1}^{k} CA^{k-t}w_t\right)v_k^T$$
$$+ 2C\mathcal{A}\left(\left(\sum_{t=-1}^{k} A^{k-t}w_t\right) w_k^T\right) C^T - C\mathcal{A}\left(w_kw_k^T\right) C^T,$$

and the initial condition $\mathcal{N}_{-1} = \mathcal{A}\left(x_{-1}x_{-1}^T\right)$, $\mathcal{M}_{-1} = CA\mathcal{N}_{-1}A^TC^T$.

One can then prove that $\mathbb{E}(\Pi_k|\mathcal{F}_{k-1}) = \mathcal{W}$, $\mathbb{E}\|\Pi_k - \mathcal{W}\|^2 \sim O(1)$. Hence, $\mathcal{M}_k - k\mathcal{W}$ is a martingale and $\mathcal{M}_k/k - \mathcal{W} \sim \mathcal{C}(-0.5)$ by Lemma 4. On the other hand, for $\mathcal{N}_k$, since $A\otimes A$ is stable, $\mathcal{N}_k \sim \mathcal{C}(0)$ by Lemma 3, which proves $\frac{1}{k+1}\sum_{t=0}^{k} \vartheta_t\vartheta_t^T - \mathcal{W} \sim \mathcal{C}(-0.5)$. ■

By Lemma 3, $\mathcal{W}_k - \mathcal{W} \sim \mathcal{C}(\max\{-0.5, -\gamma, -2\gamma\}) = \mathcal{C}(-\gamma)$. By Lemma 3.3, one can prove that $\mathcal{P}_k - \mathcal{P}$, $\mathcal{X}_k - \mathcal{X}$ are all of the class $\mathcal{C}(-\gamma)$, as they are differentiable functions of $\lambda_{k,i}$, $\Omega_{k,i}$ and $\mathcal{W}_k$. Therefore, $U_{k,*} - U_* \sim \mathcal{C}(-\gamma)$ since $U_{k,*}$ is a differentiable function of $\mathcal{P}_k$ and $\mathcal{X}_k$ at a neighborhood of $\mathcal{P}$ and $\mathcal{X}$ (see [26]).

Hence, one can prove that $\mathcal{U}_k - \mathcal{U} \sim \mathcal{C}(-\gamma)$, as $\mathcal{U}_k$ is a differentiable function of $\lambda_{k,i}$, $\Omega_{k,i}$ and $U_{k,*}$.

Finally we prove that $\hat{g}_k - g_k \sim \mathcal{C}(-\gamma)$ due to Lemma 3.1.

References

[1] E. A. Lee and S. A. Seshia, *Introduction to embedded systems: A cyber-physical systems approach*. MIT Press, 2016.

[2] A. Humayed, J. Lin, F. Li, and B. Luo, "Cyber-physical systems security—a survey," *IEEE Internet of Things Journal*, vol. 4, no. 6, pp. 1802–1831, 2017.

[3] H. Sandberg, S. Amin, and K. H. Johansson, "Cyberphysical security in networked control systems: An introduction to the issue," *IEEE Control Systems*, vol. 35, no. 1, pp. 20–23, 2015.

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

[5] U. P. D. Ani, H. He, and A. Tiwari, "Review of cybersecurity issues in industrial critical infrastructure: manufacturing in perspective," *Journal of Cyber Security Technology*, vol. 1, no. 1, pp. 32–74, 2017.

[6] A. Cardenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, S. Sastry *et al.*, "Challenges for securing cyber physical systems," in *Workshop on future directions in cyber-physical systems security*, vol. 5, 2009.

[7] C. Neuman, "Challenges in security for cyber-physical systems," in *DHS workshop on future directions in cyber-physical systems security*. Citeseer, 2009, pp. 22–24.

[8] D. Gollmann and M. Krotofil, "Cyber-physical systems security," in *The New Codebreakers*. Springer, 2016, pp. 195–204.

[9] R. Mitchell and I.-R. Chen, "A hierarchical performance model for intrusion detection in cyber-physical systems," in *Wireless Communications and Networking Conference (WCNC), 2011 IEEE*. IEEE, 2011, pp. 2095–2100.

[10] R. Mitchell and I.-R. Chen, "A survey of intrusion detection techniques for cyber-physical systems," *ACM Computing Surveys (CSUR)*, vol. 46, no. 4, p. 55, 2014.

[11] C. Kwon, W. Liu, and I. Hwang, "Security analysis for cyber-physical systems against stealthy deception attacks," in *American Control Conference (ACC), 2013*. IEEE, 2013, pp. 3344–3349.

[12] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[13] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*. IEEE, 2009, pp. 911–918.

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

[15] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on scada systems," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, 2014.

[16] B. Satchidanandan and P. R. Kumar, "Dynamic Watermarking: Active Defense of Networked Cyber–Physical Systems," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 219–240, feb 2017.

[17] 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, jul 2017.

[18] Fei Miao, M. Pajic, and G. J. Pappas, "Stochastic game approach for replay attack detection," in *52nd IEEE Conference on Decision and Control*. IEEE, dec 2013, pp. 1854–1859.

[19] A. Hoehn and P. Zhang, "Detection of replay attacks in cyber-physical systems," in *American Control Conference (ACC), 2016*. IEEE, 2016, pp. 290–295.

[20] H. Liu, J. Yan, Y. Mo, and K. H. Johansson, "An on-line design of physical watermarks," in *2018 IEEE Conference on Decision and Control (CDC)*, Dec 2018, pp. 440–445.

[21] C.-T. Chen, *Linear system theory and design*. Oxford University Press, Inc., 1998.

[22] L. L. Scharf and C. Demeure, *Statistical signal processing: detection, estimation, and time series analysis*. Addison-Wesley Reading, MA, 1991, vol. 63.

[23] J. J. Downs and E. F. Vogel, "A plant-wide industrial process control problem," *Computers & chemical engineering*, vol. 17, no. 3, pp. 245–255, 1993.

[24] N. L. Ricker, "Model predictive control of a continuous, nonlinear, two-phase reactor," *Journal of Process Control*, vol. 3, no. 2, pp. 109–123, 1993.

[25] Y. S. Chow, "On a Strong Law of Large Numbers for Martingales," *The Annals of Mathematical Statistics*, vol. 38, no. 2, pp. 610–610, apr 1967.

[26] D. R. Brillinger, "The analyticity of the roots of a polynomial as functions of the coefficients," *Mathematics Magazine*, vol. 39, no. 3, pp. 145–147, 1966.

[27] H. Liu, Y. Mo, J. Yan, L. Xie, and K. H. Johansson, "An On-line Approach to Physical Watermark Design," *arXiv e-prints*, p. arXiv:1911.01868, Nov 2019.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.