2018 IEEE Conference on Decision and Control (CDC)
Miami Beach, FL, USA, Dec. 17-19, 2018

An On-line Design of Physical Watermarks

Hanxiao Liu, Jiaqi Yan, Yilin Mo, Karl Henrik Johansson

***Abstract*— This paper considers the problem of designing physical watermark signals to protect a control system against replay attacks. We first introduce the replay attack model, where an adversary replays the previous sensory data in order to fool the controller to believe the system is still operating normally. The physical watermarking scheme, which leverages a random control input as a watermark to detect the replay attack is introduced. The optimal watermark signal design problem is then proposed as an optimization problem, which achieves the optimal trade-off between the control performance and attack detection performance. For the system with unknown parameters, we provide a procedure to asymptotically derive the optimal watermarking signal. Numerical examples are provided to illustrate the effectiveness of the proposed strategy.**

I. Introduction

Cyber-Physical Systems (CPS) are defined as the system where "*physical and software components are deeply intertwined, each operating on different spatial and temporal scales, exhibiting multiple and distinct behavioral modalities, and interacting with each other in a myriad of ways that change with context*" [1]. It plays a vital role in a large variety of fields, such as manufacturing, health care, transportation, military and infrastructure construction. Due to the wide applications and critical functions of the CPS, increasing importance has been attached to the security of CPS [2], [3]. A successful attack can jeopardize critical infrastructure and people's lives and properties, even threaten national security. Therefore, the design of secure CPS and defense mechanisms becomes crucial to ensuring proper operation of CPS [4].

However, CPS security faces a wide variety of challenges. Cardenas *et al.* [5] discuss three main challenges and identify the unique properties of CPS security when compared with traditional IT security. Taylor and Sharif [6] review the difficulties of guaranteeing the critical infrastructure systems and industrial control systems. The research community has made significant efforts in false data injection, failure and anomaly detection to enhance CPS security in recent years. Manandhar and Cao [7] propose a robust security framework for the smart-grid system using the χ^2 detector and Euclidean detector. The fault detection problem for linear time-invariant discrete-time systems with disturbance is analyzed in [8].

In this paper, we consider the problem of detecting replay attack, which is motivated by the Stuxnet malware. In [9], [10], [11], 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 the presence of the attack, by injecting a random control signal, namely watermark signal, into the control system. However, the watermark signal will deteriorate the control performance, and therefore it is important to find the optimal trade-off between the control performance loss and the detection performance, which can be casted as an optimization problem. Similar "watermarking" schemes are also proposed in the literature [12], [13], [14].

It is worth noticing that in the majority of the aforementioned researches, the precise knowledge of the system parameters is assumed in order to design the watermarking signal. However, acquiring the parameters may be troublesome and costly. Hence, it is beneficial for the system to "learn" the parameters during its operation and automatically design the watermarking signal in real-time. Motivated by this idea, in this paper, we propose a "on-line learning mechanism" to infer the system parameters. The physical watermark that asymptotically converges to the optimal one is further developed.

The rest of paper is organized as follows. Section II formulates the problem by introducing the system as well as the attack model. The physical watermarking scheme is introduced in Section III. In Section IV, we present an on-line "learning" scheme based on the input and output data to infer the parameters of the system and design the watermark signal based on the estimated parameters. We further prove the almost sure convergence of the watermarking signal to the optimal one. In Section V, numerical example is provided to verify the effectiveness of the proposed technique. Concluding remarks are given in Section VI.

Notations: $\|A\|_F$ is the Frobenius norm of an $m\times n$ matrix A defined as $\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{i,j}^2}$, where $A_{i,j}$ is the ith row, jth column element of the matrix A. $A\otimes B$ is the Kronecker product of matrix A and B. $A>0$ denotes that the matrix A is positive definite. A^T denotes the transpose of matrix A.

II. Problem Formulation

In this section, we introduce the system as well as the attack model, which will be used for the remaining of the paper.

We consider a linear time-invariant system described by

H. Liu, J. Yan and Y. Mo are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. Email: {hanxiao001, jyan004, ylmo}@ntu.edu.sg.

K.H. Johansson is with the ACCESS and the Department of Automatic Control, the School of Electrical Engineering, 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 following equations:

$$x_{k+1} = Ax_k + w_k, \tag{1}$$
$$y_k = Cx_k + v_k, \tag{2}$$

where $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}^m$ are the state vector and the sensor's measurement, respectively; $w_k \in \mathbb{R}^n$ is the zero mean Gaussian process noise with covariance $Q > 0$, and $v_k \in \mathbb{R}^n$ is the zero mean Gaussian measurement noise with covariance $R > 0$. We suppose that $w_0, w_1, \dots$ and $v_0, v_1, \dots$ are independent of each other. We further assume that x_0 is a zero mean Gaussian random vector independent of the process noise and the measurement noise, with covariance Σ.

We further make the following assumptions regarding the system:

Assumption 1: The A matrix is strictly stable. Furthermore, (A, C) is observable.

Notice that the observability assumption is without loss of generality as we can perform a Kalman decomposition and only work with the observable subspace.

Since CPS usually operates for an extended period of time, we assume that the system is already in the steady state, i.e., Σ satisfies:

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

Next we introduce the replay attack model. The adversary is assumed to have the following capabilities:

1) The attacker has access to all the real-time sensory data. In other words, it knows $y_0, \dots, y_k$ at time k.
2) The attacker can modify the true 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 $k_1 + T_p$, where T_p is large enough to guarantee that the attacker can replay the sequence for an extended period of time during the attack.
2) The attacker manipulates the sensor measurements y_k starting from time k_2 to the recorded signals, i.e.,

$$y'_k = y_{k-\Delta k}, \forall\ k_2 \le k \le (k_2 + T_p),$$

where $\Delta k = k_2 - k_1$.

It is worth noticing that since the system is already in the steady state, the statistics of replayed y'_k will be exactly as the same as that of the real data y_k. As a result, for a large class of linear systems, the replayed signal and the real one become indistinguishable after a short transient time period. For more detailed discussion, please refer to [9].

III. Physical Watermarking Scheme

This section is devoted to the detection of replay attack via physical watermarking. The main idea of physical watermarking is to inject a random noise ϕ_k, which is called the watermark signal, to excite the system and check if the system responds to the watermark signal in accordance to the dynamical model of the system. To be specific, we assume that the system equation (1) is modified to be

$$x_{k+1} = Ax_k + B\phi_k + w_k, \tag{4}$$

where $\phi_k \in \mathbb{R}^p$ is the watermark signal applied to the system at time k, which is usually assumed to be independent and identically distributed (i.i.d.) zero mean Gaussian with variance U. We further assume that (A, B) is controllable.

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

$$y_k = \sum_{t=0}^{k-1} CA^t B\phi_{k-1-t} + \sum_{t=0}^{k-1} CA^t w_{k-1-t} + v_k + CA^k x_0.$$

For simplicity, we define

$$\gamma_k \triangleq \sum_{t=0}^{k} CA^t B\phi_{k-t},$$
$$\vartheta_k \triangleq \sum_{t=0}^{k} CA^t w_{k-t} + v_{k+1} + CA^{k+1} x_0.$$

Hence, y_k can be rewritten in the following form:

$$y_k = \gamma_{k-1} + \vartheta_{k-1}. \tag{5}$$

We further define

$$H_\tau \triangleq CA^\tau B.$$

One can check that γ_{k-1} is a zero mean Gaussian whose covariance converges to $\mathcal{U}$, where

$$\mathcal{U} = \sum_{\tau=0}^{\infty} H_\tau U H_\tau^T.$$

Similarly, ϑ_k is a zero mean Gaussian noise with covariance $\mathcal{W} = C\Sigma C^T + R$, where Σ is defined in (3). As a result, given $\phi_0, \dots, \phi_{k-1}$, the conditional distribution of y_k converges to a Gaussian distribution with mean γ_{k-1} and covariance $\mathcal{W}$.

For the scenario where replay attack is present, the replayed y'_k can be written as

$$\begin{aligned} y'_k &= y_{k-\Delta k} \\ &= \gamma_{k-1-\Delta k} + \vartheta_{k-1-\Delta k}. \end{aligned}$$

Since Δk is unknown to the system operator, it is safe to assume that given $\phi_0, \dots, \phi_{k-1}$, y'_k is zero mean Gaussian with covariance $\mathcal{U} + \mathcal{W}$.

As a result, we can design a detector to differentiate the distribution of y_k under the following two hypotheses:
H_0: y_k follows a Gaussian distribution $\mathcal{N}_0 = \mathcal{N}(\gamma_{k-1}, \mathcal{W})$.
H_1: y_k follows a Gaussian distribution $\mathcal{N}_1 = \mathcal{N}(0, \mathcal{U} + \mathcal{W})$.

By the Neyman-Pearson lemma [15], the Neyman-Pearson detector for hypothesis H_0 versus hypothesis H_1 takes the following form:

Theorem 1: The Neyman-Pearson detector rejects H_0 in favor of H_1 if

$$\begin{aligned} &g(y_k, \phi_{k-1}, \phi_{k-2}, \cdots) \\ =&\left(y_k - \gamma_{k-1}\right)^T W^{-1} \left(y_k - \gamma_{k-1}\right) - y_k^T \left(W + \mathcal{U}\right)^{-1} y_k \\ \ge&\eta, \end{aligned} \tag{6}$$

where η is a predetermined threshold depending on the desired false alarm rate. Otherwise, hypothesis H_0 is accepted.

Similar to [10], the quantity $tr(\mathcal{U}\mathcal{W}^{-1})$ can be used to characterize the detection performance. In other words, increasing $tr(\mathcal{U}\mathcal{W}^{-1})$ will usually results in better detection performance. For more details, please refer to [10].

Note that although the watermark signal can enable the detection of replay attack, it also deteriorates the performance of the system. As a result, it is important to find 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 LQG metric:

$$J = \lim_{T\to+\infty} \mathbb{E}\left(\frac{1}{T}\sum_{k=0}^{T-1}\begin{bmatrix} y_k \\ \phi_k \end{bmatrix}^T X \begin{bmatrix} y_k \\ \phi_k \end{bmatrix}\right), \tag{7}$$

where

$$X = \begin{bmatrix} X_{yy} & X_{y\phi} \\ X_{\phi y} & X_{\phi\phi} \end{bmatrix} > 0.$$

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

$$J = \lim_{k\to} \operatorname{tr}\left(X \operatorname{Cov}\left(\begin{bmatrix} y_k \\ \phi_k \end{bmatrix}\right)\right) = \operatorname{tr}\left(X \begin{bmatrix} \mathcal{W}+\mathcal{U} & H_0U \\ UH_0^T & U \end{bmatrix}\right).$$

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

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

with S being a following linear function of U,

$$S = \begin{bmatrix} \mathcal{U} & H_0U \\ UH_0^T & U \end{bmatrix}.$$

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

$$\begin{aligned} U = \arg\max_{U\geq 0} \quad & \operatorname{tr}(\mathcal{U}\mathcal{W}^{-1}) \\ \text{subject to} \quad & \operatorname{tr}(XS) \leq \delta, \end{aligned} \tag{8}$$

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

An important property of the optimization problem (8) is that the optimal solution is usually a rank-1 matrix, which is formalized by the following theorem:

Theorem 2: The optimization problem (8) is equivalent to

$$\begin{aligned} U = \arg\max_{U\geq 0} \quad & \operatorname{tr}(U\mathcal{P}) \\ \text{subject to} \quad & \operatorname{tr}(U\mathcal{X}) \leq \delta, \end{aligned} \tag{9}$$

where

$$\mathcal{P} \triangleq \sum_{\tau=0}^{\infty} H_\tau^T \mathcal{W}^{-1} H_\tau, \tag{10}$$

$$\mathcal{X} \triangleq \left(\sum_{\tau=0}^{\infty} H_\tau^T X_{yy} H_\tau\right) + H_0^T X_{y\phi} + X_{\phi y} H_0 + X_{\phi\phi}. \tag{11}$$

The optimal solution (not necessarily unique) to (9) 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.

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

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

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 $\mathcal{X} > 0$.

The proof of the second part is similar to the proof of Theorem 7 in [11] and is omitted here due to space limit. ■

It is worth noticing that in order to design the optimal watermarking signal, precise knowledge of the system parameters is needed. However, acquiring the parameters may be troublesome and costly. Therefore, it is beneficial for the system to "learn" the parameters during its operation and design the watermarking signal in real time, which will be our focus in the next section.

IV. On-line "Learning" Scheme

This section is devoted to developing an on-line "learning" procedure to find the optimal watermarking signals. Throughout the section, we make the following assumptions:

1) A is diagonalizable and has distinct eigenvalues.
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 system output y_k, the dimension of the A matrix n is known, the matrix X and δ are known.

For the sake of legibility, we first introduce how to infer the necessary parameters of the system. Then we move to the design of watermark signal based on the estimated parameters. The proofs of Theorem 3, 4 and 5 are reported at the end of this section.

A. Inference on the Parameters

In this subsection, we describe our "learning" procedure. At each time k, the watermarking 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 matrix U_k is computed as a function of $y_0, \ldots, y_k, \phi_0, \ldots, \phi_{k-1}$, the procedure of which will be described in details in the next subsection.

Define Y_k and $H_{k,\tau}$ $(0 \leq \tau \leq 3n-2)$ as

$$Y_k \triangleq \frac{1}{k+1}\sum_{t=0}^{k} y_t y_t^T,\ H_{k,\tau} \triangleq \frac{1}{k+1}\sum_{t=0}^{k} y_t \phi_{t-\tau-1}^T U_{t-\tau-1}^{-1}.$$

We shall assume that $\phi_{t-\tau-1} = 0$ if $t-\tau-1 < 0$.

One can think $H_{k,\tau}$ is an estimate of H_τ and Y_k is an estimate of $\mathcal{W}+\mathcal{U}$. We first prove a theorem regarding the convergence $H_{k,\tau}$ to H_τ.

Theorem 3: Suppose that there exists positive definite matrices $\overline{M}$ and $\underline{M}$, such that the following inequality surely holds:

$$\overline{M} > U_k > \frac{1}{(k+1)^\beta}\underline{M}, \tag{12}$$

where $0 \leq \beta < 1$, then $H_{k,\tau}$ converges to H_τ almost surely.

It is worth noticing that we can only keep a record of finitely many $H_{k,\tau}$s. However, to infer matrices $\mathcal{U}, \mathcal{W}, \mathcal{P}$ and $\mathcal{X}$, we need to estimate H_τ for all $\tau \geq 0$. The following lemma provides a method to obtain H_τ from only finite parameters and its proof can be found in [16].

Lemma 1: Suppose that the matrix A has distinct eigenvalues $\lambda_1, \ldots, \lambda_n$, then there exists unique $\Omega_1, \ldots, \Omega_n$, such that

$$H_\tau = \sum_{i=1}^{n} \lambda_i^\tau \Omega_i. \tag{13}$$

By Lemma 1, we could use finitely many $H_0, \ldots, H_{3n-2}$ to estimate both λ_is and Ω_is and thus H_τ for any τ. To this end, let us consider the following optimization problem:

$$\min_{\alpha_{k,0},\ldots,\alpha_{k,n-1}} \left\| \mathcal{H}_k \left(\begin{bmatrix} \alpha_{k,0} \\ \alpha_{k,1} \\ \cdots \\ \alpha_{k,n-1} \end{bmatrix} \otimes I \right) + \begin{bmatrix} H_{k,n} \\ H_{k,n+1} \\ \cdots \\ H_{k,3n-2} \end{bmatrix} \right\|_F, \tag{14}$$

where $\mathcal{H}_k$ is a Hankel matrix defined as

$$\mathcal{H}_k \triangleq \begin{bmatrix} H_{k,0} & H_{k,1} & \ldots & H_{k,n-1} \\ H_{k,1} & H_{k,2} & \ldots & H_{k,n} \\ \vdots & \vdots & \ddots & \vdots \\ H_{k,2n-2} & H_{k,2n-1} & \ldots & H_{k,3n-3} \end{bmatrix}.$$

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

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

and

$$\begin{bmatrix} \Omega_{k,1} \\ \vdots \\ \Omega_{k,n} \end{bmatrix} = \arg\max_{\Omega_{k,i}} \left\| (V_k \otimes I_m) \begin{bmatrix} \Omega_{k,1} \\ \vdots \\ \Omega_{k,n} \end{bmatrix} - \begin{bmatrix} H_{k,0} \\ \cdots \\ H_{k,3n-2} \end{bmatrix} \right\|.$$

The following theorem further establishes the convergence of $\lambda_{k,i}$ (and $\Omega_{k,i}$) to λ_i (and Ω_i):

Theorem 4: Suppose that A has distinct eigenvalues. If $H_{k,\tau}$ converges to H_τ for $0 \leq \tau \leq 3n-2$, then $\lambda_{k,i}$ converges λ_i and $\Omega_{k,i}$ converges to Ω_i.

Then let us define $\mathcal{U}_{k,ij}$, which satisfies the following recursive equation:

$$\mathcal{U}_{k+1,ij} = \lambda_{k,i}\lambda_{k,j}\mathcal{U}_{k,ij} + \Omega_i U_k \Omega_j^T,$$

and

$$\mathcal{U}_k \triangleq \sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{U}_{k,ij}.$$

Furthermore, define

$$\mathcal{W}_k = Y_k - \frac{1}{k+1}\sum_{t=0}^{k} \mathcal{U}_t.$$

The following theorem establishes the convergence of $\mathcal{W}_k$:

Theorem 5: Suppose that (12) holds, then $\mathcal{W}_k$ converges to $\mathcal{W}$ almost surely.

Let us further define

$$\mathcal{P}_k = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{1-\lambda_{k,i}\lambda_{k,j}} \Omega_{k,i}^T \mathcal{W}_k^{-1} \Omega_{k,j},$$

and

$$\begin{aligned}\mathcal{X}_k = &\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{1}{1-\lambda_{k,i}\lambda_{k,j}} \Omega_{k,i}^T X_{yy} \Omega_{k,j} \\ &+ \sum_{i=1}^{n} \Omega_i^T X_{y\phi} + X_{\phi y}\sum_{i=1}^{n}\Omega_i + X_{\phi\phi}.\end{aligned}$$

By the convergence of $H_{k,\tau}$, $\mathcal{W}_k$, $\lambda_{k,i}$ and $\Omega_{k,i}$, it is easy to prove that $\mathcal{P}_k$ and $\mathcal{X}_k$ converges to $\mathcal{P}$ and $\mathcal{X}$ almost surely. As a result, we have successfully estimated all the parameters necessary to design the watermarking signal, with the only assumption being (12).

B. Watermarking Signal Design

U_k is updated as

$$U_{k+1} = U_{k,*} + \frac{\delta}{(k+1)^\beta} I, \tag{15}$$

where δ is defined in (8) and $U_{k,*}$ is the solution of the following optimization problem

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

$0 \leq \beta < 1$. The following theorem establishes the boundedness and convergence of U_k.

Theorem 6: U_k is bounded by

$$\delta(X_{\phi\phi} - X_{\phi y}X_{yy}^{-1}X_{y\phi})^{-1} \geq U_k \geq \delta(k+1)^{-\beta} I \tag{16}$$

Furthermore, if $\mathcal{P}_k$ converges to $\mathcal{P}$ and $\mathcal{X}_k$ converges to $\mathcal{X}$, then

$$\lim_{k\to\infty} U_k = U,$$

where U is the solution of (14).

Proof: Notice that

$$\mathcal{X}_k \geq \left(\sum_{i=1}^{n}\Omega_i\right)^T X_{yy}\left(\sum_{i=1}^{n}\Omega_i\right) + \sum_{i=1}^{n}\Omega_i^T X_{y\phi} + X_{\phi y}\sum_{i=1}^{n}\Omega_i + X_{\phi\phi}.$$

Hence, $\mathcal{X}_k \geq X_{\phi\phi} - X_{\phi y}X_{yy}^{-1}X_{y\phi}$, which implies that

$$\mathrm{tr}(U_{k,*}\left(X_{\phi\phi} - X_{\phi y}X_{yy}^{-1}X_{y\phi}\right)) \leq \delta. \tag{17}$$

Notice that if for a positive semidefinite X with $\mathrm{tr}(X) \leq \delta$, then $X \leq \delta I$. Hence, (17) implies that

$$U_{k,*} \leq \delta\left(X_{\phi\phi} - X_{\phi y}X_{yy}^{-1}X_{y\phi}\right)^{-1},$$

which proves the first inequality in (16). The second inequality can be easily proved by (15).

The convergence can be proved by noticing that $U_{k,*}$ is a continuous function of $\mathcal{P}_k$, $\mathcal{X}_k$ at a neighborhood of $\mathcal{P}, \mathcal{X}$. The detailed proof is omitted due to space limit. ∎

Now we can establish that U_k converges to the optimal U. Notice that there is no circular logic in our proof, as (16) holds regardless of the inferred value Y_k and $H_{k,\tau}$. Therefore, the convergence of $\mathcal{X}_k$ and $\mathcal{P}_k$ is guaranteed by Theorem 3, 4 and 5, which further implies the convergence of U_k.

C. Proofs of Theorem 3, 4 and 5

1) Proof of Theorem 3: We only prove for the case where $\tau = 0$. The $\tau > 0$ case can be proved following similar arguments and the details are omitted due to space constraints. Before proving theorem 3, the following lemmas are needed and their proofs can be found in [16].

Lemma 2: Suppose that $\omega, \upsilon, \varsigma, \xi$ are four jointly Gaussian random vectors with zero mean and proper dimensions. The following equations are true:

$$\mathbb{E}\left[\omega\upsilon^T\varsigma\xi^T\right] = \mathbb{E}\left[\omega\xi^T\right]\mathbb{E}\left[\upsilon^T\varsigma\right] + \mathbb{E}\left[\omega\varsigma^T\right]\mathbb{E}\left[\upsilon\xi^T\right] + \mathbb{E}\left[\omega\upsilon^T\right]\mathbb{E}\left[\varsigma\xi^T\right],$$

$$\mathbb{E}\left[\upsilon^T\varsigma\xi^T\right] = 0.$$

Lemma 3: If $\Upsilon_n = \Pi_0 + \cdots + \Pi_n$ be a martingale such that

$$\sum_{k=0}^{\infty}\frac{\mathbb{E}\left\|\Pi_k\right\|_F^2}{(k+1)^2} < \infty,$$

where $\Pi_k (k = 0, \cdots, n)$ and Υ_n are all $m \times l$ matrices, then

$$\lim_{n\to\infty}\frac{\Upsilon_n}{n+1} = \mathbf{0} \text{ almost surely.}$$

Proof: [Proof of Theorem 3] We only provide an outline of the proof. The detailed proof can be found in [16]. Define the filtration $\mathcal{F}_k$ to be the σ-algebra which is generated by the following random variables $\{x_0, \phi_0, \ldots, \phi_{k-1}, w_0, \ldots, w_{k-1}, v_0, \ldots, v_k\}$. It is easy to see that both U_k and y_k are measurable in the σ-algebra $\mathcal{F}_k$. Let us further define

$$\mathcal{S}_k = \sum_{t=0}^{k}(y_t\phi_{t-1}^T U_{t-1}^{-1} - H_0),$$

where $\phi_{k-1} = 0$ if $k < 1$. The proof is divided into steps.

First, one can prove that S_k is a martingale with respect to the filtration $\{\mathcal{F}_k\}$, i.e.,

$$\mathbb{E}(\mathcal{S}_{k+1}|\mathcal{F}_k) = \mathcal{S}_k, \tag{18}$$

or in other words,

$$\mathbb{E}(y_{k+1}\phi_k^T U_k^{-1}|\mathcal{F}_k) = H_0.$$

Next we need to prove that

$$\sum_{k=0}^{\infty}\frac{\mathbb{E}\left\|y_{k+1}\phi_k^T U_k^{-1} - H_0\right\|_F^2}{(k+1)^2} < \infty. \tag{19}$$

Let us consider

$$\begin{aligned}&\left[y_{k+1}\phi_k^T U_k^{-1} - H_0\right]\left[y_{k+1}\phi_k^T U_k^{-1} - H_0\right]^T\\ =&y_{k+1}\phi_k^T U_k^{-2}\phi_k y_{k+1}^T - H_0 U_k^{-1}\phi_k y_{k+1}^T\\ &- y_{k+1}\phi_k^T U_k^{-2}H_0^T + H_0 H_0^T,\end{aligned}$$

Now by Lemma 2, we can prove that

$$\begin{aligned}&\mathbb{E}\left(\left[y_{k+1}\phi_k^T U_k^{-1} - H_0\right]\left[y_{k+1}\phi_k^T U_k^{-1} - H_0\right]^T|\mathcal{F}_k\right)\\ =&H_0 U_k H_0^T\,\mathrm{tr}(U_k^{-1}) + \mathrm{tr}(U_k^{-1})\psi_{k+1}\psi_{k+1}^T + \mathrm{tr}(U_k^{-1})R\\ &+ H_0 H_0^T.\end{aligned}$$

Now if $\overline{M} \geq U_k \geq \underline{M}/(k+1)^\beta$, we can conclude that

$$\begin{aligned}&\mathbb{E}\left(\left\|y_{k+1}\phi_k^T U_k^{-1} - H_0\right\|_F^2\right)\\ =&\mathrm{tr}\left(\mathbb{E}\left(\left[y_{k+1}\phi_k^T U_k^{-1} - H_0\right]\left[y_{k+1}\phi_k^T U_k^{-1} - H_0\right]^T\right)\right)\\ =&O\left((k+1)^\beta\right).\end{aligned}$$

Since $\beta < 1$, according to the convergence condition of infinite series, we know that the infinite sum on LHS of (19) is bounded.

Therefore, by Lemma 3,

$$\lim_{k\to\infty}\frac{S_k}{k+1} = \mathbf{0} \text{ almost surely,}$$

which proves that $H_{k,0}$ converges to H_0 almost surely. ∎

2) Proof of Theorem 4: Before proving Theorem 4, we need the following lemma, whose proof can be found in [16].

Lemma 4: Suppose that the vector φ is the solution of the optimization problem

$$\varphi = \arg\min_{\varphi}\|A(\theta)\varphi - b(\theta)\|_2,$$

where $A(\theta)$ and $b(\theta)$ are continuous functions of θ. If $A(\theta_0)$ is of full column rank at θ_0, then φ is unique and a continuous function of θ in a neighborhood of θ_0.

The proof of Theorem 4 can be proved by Lemma 1 and Lemma 4. The details can be found in [16].

3) Proof of Theorem 5: Before proving the theorem, we need the following lemma, whose proof can be found in [16].

Lemma 5: Suppose that ρ_k converges to ρ, where $|\rho| < 1$. Furthermore, assume that $\lim_{k\to\infty} a'_k - a_k = 0$, where a_k is a bounded sequence. Then we have

$$\lim_{k\to\infty} b'_k - b_k = 0,$$

where b_k and b'_k satisfy the following recursive equation:

$$b_{k+1} = \rho b_k + a_k,\ b'_{k+1} = \rho_k b'_k + a'_k,$$

with initial condition $b_{-1} = b'_{-1} = 0$.

The proof of Theorem 5 can be proved by Lemma 1, Lemma 5 and Theorem 6 in [17]. The detailed proof can be found in [16].

V. Simulation Result

In this section, the performance of the proposed learning procedure is evaluated. We choose $n = m = p = 2$ and A, B, C are all randomly generated, with A stable.

Without loss of generality, it is assumed that X in (7), the covariance matrices Q and R are equal to the identity matrix with proper dimensions. We assume that δ in (9) is equal to 5 and $\beta = 1/3$. Figure 1 shows $\|U_k - U\|_F/\|U\|$ v.s. time k, where U is the solution of the optimization problem of (8), and U_k, generated through updating equation (15), is the estimation of U.

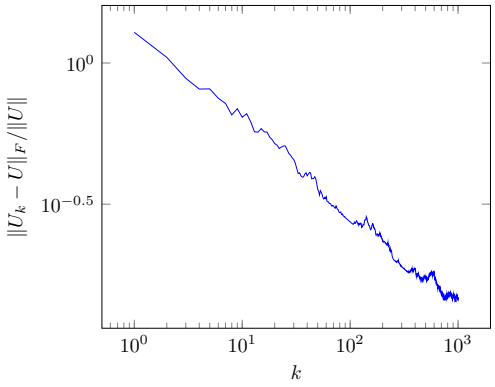

Fig. 1: $\|U_k - U\|_F/\|U\|$ versus k.

From Figure 1, it can be seen that U_k converges to the optimal U as time goes to infinity. Furthermore, the convergence follows a power law, i.e., $\|U_k - U\|_F = O(k^{-\epsilon})$. We plan to investigate the rate of the convergence in our future work.

VI. Conclusion

In this paper, the detection problem of replay attack via "physical watermarking" with known system parameters is proposed to achieve the desired trade-off between the detection performance and control performance loss. Then we provide an on-line "learning" technique for determining the optimal watermarking signals without the knowledge of system parameters. The simulation is carried out to verify the effectiveness of the proposed technique.

References

[1] N. S. Foundation, "Cyber physical systems nsf10515," 2010. [Online]. Available: http://www.nsf.gov/pubs/2010/nsf10515/nsf10515.htm

[2] A. Humayed, J. Lin, F. Li, and B. Luo, "Cyber-physical systems securitya 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] 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.

[5] 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.

[6] J. M. Taylor and H. R. Sharif, "Security challenges and methods for protecting critical infrastructure cyber-physical systems," in *Selected Topics in Mobile and Wireless Networking (MoWNeT), 2017 International Conference on*. IEEE, 2017, pp. 1–6.

[7] K. Manandhar, X. Cao, F. Hu, and Y. Liu, "Detection of faults and attacks including false data injection attack in smart grid using kalman filter," *IEEE transactions on control of network systems*, vol. 1, no. 4, pp. 370–379, 2014.

[8] H. Wang and G.-H. Yang, "A finite frequency domain approach to fault detection for linear discrete-time systems," *International Journal of Control*, vol. 81, no. 7, pp. 1162–1171, 2008.

[9] 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.

[10] 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.

[11] 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.

[12] A. Khazraei, H. Kebriaei, and F. R. Salmasi, "A new watermarking approach for replay attack detection in lqg systems," in *Decision and Control (CDC), 2017 IEEE 56th Annual Conference on*. IEEE, 2017, pp. 5143–5148.

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

[14] A. Khazraei, H. Kebriaei, and F. R. Salmasi, "Replay attack detection in a multi agent system using stability analysis and loss effective watermarking," in *American Control Conference (ACC), 2017*. IEEE, 2017, pp. 4778–4783.

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

[16] H. Liu, J. Yan, Y. Mo, and K. H. Johansson, "An on-line design of physical watermarks," 2018. [Online]. Available: https://arxiv.org/abs/1809.05299

[17] R. Lyons, "Strong laws of large numbers for weakly correlated random variables." *The Michigan Mathematical Journal*, vol. 35, no. 3, pp. 353–359, 1988.