Active Detection against Replay Attack: A Survey on Watermark Design For Cyber-Physical Systems

Hanxiao Liu, Yilin Mo and Karl Henrik Johansson

Abstract Watermarking is a technique that embeds digital information, "watermark", in a carrier signal to identify ownership of the signal or verify the authenticity or integrity of the carrier signal. It has been widely employed in the fields of image and signal processing. In this chapter, we survey some recent physical watermark design approaches for Cyber-Physical Systems (CPS). We focus on how to design physical watermarking to actively detect cyber attacks, especially replay attacks, thereby securing the CPS. First, the system as well as the attack model are introduced. A basic physical watermarking scheme, which leverages a random noise as a watermark to detect the attack, is discussed. The optimal watermark signal is designed to achieve a trade-off between control performance and intrusion detection. Based on this scheme, several extensions are also presented, such as watermarks generated by a Hidden-Markov Model and on-line data-based watermark generation. These schemes all use an additive watermarking signal. A multiplicative watermark scheme is also presented. The chapter is concluded with a discussion on some open problems on watermark design.

H. Liu
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, and School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden
e-mail: hanxiao001@ntu.edu.sg

Y. Mo
Department of Automation and BNRist, Tsinghua University, China
e-mail: ylmo@tsinghua.edu.cn

K. H. Johansson
School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden
e-mail: kallej@kth.se

1 Introduction

Cyber-Physical Systems (CPS) integrate computational elements and physical processes closely. They are playing a more and more critical role in a large variety of infrastructures, such as transportation, power grid, defense and environment. Most of them are of great importance to the operation of society. Any successful cyber-physical attack may bring huge damages to critical infrastructure, human lives and properties, and even threaten the national security. Maroochy water breach in 2000 [1], Stuxnet malware in 2010 [2], Ukraine power outage in 2015 [3], Venezuela blackouts in 2019 [4] and other security incidents, motivate us to pay more attention to CPS security.

Recent years have witnessed more and more research regarding how to design watermarking signals to secure CPS. Watermarking is a technique that embeds digital information, a watermark, in a carrier signal to identify ownership of the signal or verify the authenticity or integrity of the carrier signal. It has been widely employed in the fields of image and signal processing. One important application of this technique is to trace illegally copied movies where a watermark is used to determine the owner of the original movie [5, 6].

In [7, 8], a physical watermarking scheme is proposed for control systems. In this scheme, if the system is operating normally, then the effect of the carefully designed watermark signal is present in the sensor measurements. However, if the system is under attack, its effect cannot be detected. Actually, it could be considered as an active defense scheme.

Mo and Sinopoli [7] investigate the problem of the detection of replay attacks and first propose the technique of introducing an authentication signal which is called physical watermark signal later. This approach enables the detection of replay attacks where an adversary can read and modify all sensor data as well as inject a malicious input into the system. Different from false data injection attacks, this type of attack does not need knowledge of the system model to generate stealthy outputs and only replays the recorded sensor measurements to the operator, which leads to that the replayed data and the real data share exactly the same statistics and for which replay attacks cannot be detected efficiently. By injecting a random control signal, the watermark signal, into the control system, it is possible to secure the system.

The authors of [8, 9] further extend the results of [7] by providing a more general physical authentication scheme to detect the replay attacks. 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. Furthermore, Mo *et al.* [8] also characterize the relationship among the control performance loss, detection rate and the strength of the Gaussian authentication input.

The term physical watermarking is first proposed in [5] to authenticate the correct operation of CPS. As a generalization of [7, 8, 9], the technique of designing the optimal watermark signal is to maximize the expected Kullback-Leibler divergence between the distributions of the compromised and the healthy residue signals, while

guaranteeing a certain maximal control performance loss. The optimization problem is separated into two steps where the optimal direction of the signal for each frequency is first computed and then all possible frequencies are considered to find the optimal watermark signal.

The watermarking approach proposed in [10] is based on an additive watermark signal generated by a dynamical system. Conditions on the parameters of the watermark signal are obtained which ensures that the residue signal of the system under attack is unstable and the attack can be detected. An optimization problem is proposed to give a loss-effective watermark signal with a certain amount of detection rate by adjusting the design parameters. A similar problem is studied for multi-agent systems in [11].

The problem of physical watermark design under packet drops at the control input is analyzed in [12]. It is interesting that Bernoulli packet drops can obtain better detection performance compared with a purely Gaussian watermarking signal. Consequently, a Bernoulli-Gaussian watermark, which incorporates both an additive Gaussian input and a Bernoulli drop process, is jointly designed to achieve the trade-off between detection performance and control performance. The effect of the proposed watermark on closed-loop performance and detection performance is analyzed.

Satchidanandan and Kumar [13] provide a comprehensive procedure for dynamic watermarking. It suggests a private excitation signals on the control input which can be traced in the system to enable the detection of attacks. Such an active defense technique is used to secure CPS that include single-input-signal output (SISO) systems with Gaussian noise, SISO auto-regressive systems with exogenous Gaussian noise, the SISO autoregressive-moving average systems with exogenous terms, SISO systems with partial observations, multi-input-multi-output systems with Gaussian noise and extension to non-Gaussian systems. In [14], they propose necessary and sufficient conditions that the statistics of the watermark needs to satisfy in order to achieve security-guaranteeing.

It is worth noticing that in all research discussed above, 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 in practice. Motivated by this, [15] proposes an algorithm that can simultaneously generate the watermarking signal and infer the system parameters to enable the detection of attacks with unknown system parameters. It is proved that the proposed algorithm converges to the optimal one almost surely.

In [16, 17], Rubio-Hernán *et al.* define cyber adversaries and cyber-physical adversaries and point out that the detection schemes proposed by Mo and Sinopoli [7] and Mo *et al.* [5] fails to detect an attack from the latter. Besides, a multi-watermark-based detection scheme is proposed to overcome the limitation. Furthermore, in [18], a periodic and intermittent event-triggered control watermark detector is presented. The new detector strategy integrates local controllers with remote controller. It is proved that the new detector scheme can detect three adversary models defined in their work.

Although it is proved that the introduction of watermark signals enables the detection of certain replay attacks, it degrades the control performance since the control input is not optimal. Considering the loss of control performance, Fang *et al.* [19] formulate a novel attack model for the replay attack. On the basis of this model, a periodic watermarking strategy is investigated. An approximated detection performance is obtained by using the proposed periodic strategy.

Different from the additive watermarking schemes, a multiplicative sensor watermarking is proposed in [20]. In this scheme, each sensor output is watermarked. The corresponding watermark remover is employed to reconstruct the real sensor measurement from the received watermarked data. This scheme does not degrade the control performance in the absence of attacks and it could be designed independently of the design of the controller and anomaly detector. Furthermore, it also enables the isolation and identification of the replay attack. A similar scheme is applied to detect cyber sensor routing attacks [21] and false data injection attacks [22]. The physical sensor re-routing attack and the cyber measurement re-routing attack are considered in [21] and corresponding detectability and isolability of these two attacks are analyzed. In [22], Teixeira and Ferrari show how to design the watermarking filters to enable the detection of stealthy false-data injection attacks and a novel technique is proposed to solve the limitation of single-output systems.

The rest of chapter is organized as follows. Section 2 formulates the problem by introducing the system model as well as attacks model. A basic physical watermark scheme is introduced in Section 3. The optimal watermark signal is designed to achieve a trade-off between control performance and intrusion detection. In Section 4, several extensions are also presented, such as watermarks generated by a Hidden-Markov Model, on-line data-based watermark generation and a multiplicative watermark scheme. Conclusions and a discussion on some open problems on watermark design are provided in Section 5.

Notations: For an $m \times n$ matrix A, $A > 0$ ($A \geq 0$) indicates that A is positive definite (positive semidefinite), A^+ denotes the pseudo-inverse of A, and $\|A\|$ is the spectral norm of A, which is its largest singular value. For two matrices A and B, $A \otimes B$ is their Kronecker product. The notation $\mathrm{sym}(X) \triangleq \frac{X+X^T}{2}$ represents the symmetric part of a matrix X. The real part of Z is denoted by $\mathfrak{R}(Z)$.

2 Problem Setup

In this section, we setup the problem by introducing a system model as well as an attack model.

2.1 System Description

Let us consider a linear time-invariant (LTI) system described by the following equations:

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

where $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}^m$ are the state vector and the sensor's measurement, respectively, $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^n$ are process and measurement noise, respectively. It is assumed that the initial state x_0, w_k and v_k are independent Gaussian random variables, and $x_0 \sim \mathcal{N}(\bar{x_0}, \Sigma)$, $w_k \sim \mathcal{N}(0, Q)$, $v_k \sim \mathcal{N}(0, R)$. It is also assumed that (A, B) is stablizable and (A, C) is detectable.

Here, we assume that the objective of the system operator is to derive an optimal solution to minimize the following linear-quadratic-Gaussian (LQG) cost:

$$J \triangleq \lim_{T\to\infty} \mathbb{E}\frac{1}{T}\left[\sum_{k=0}^{T-1}\left(x_k^T W x_k + u_k^T V u_k\right)\right], \tag{3}$$

where W, V are positive definite matrices and u_k is measurable with respect to previous observations. Due to the separation principle, the optimal solution of (3) combines Kalman filter and LQG controller. The optimal state estimate $\hat{x}_k$ is given by Kalman filter as follows:

$$\begin{aligned}
&\hat{x}_{0|-1} = \bar{x}_0, P_{0|-1} = \Sigma,\\
&\hat{x}_{k+1|k} = A\hat{x}_{k|k} + Bu_k, P_{k+1|k} = AP_{k|k}A^T + Q,\\
&K_k = P_{k|k-1}C^T(CP_{k|k-1}C^T + R)^{-1},\\
&\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(y_k - C\hat{x}_{k|k-1}), P_{k|k} = P_{k|k-1} - K_kCP_{k|k-1}.
\end{aligned}$$

It is well known that the gain K_k converges to a fixed gain since the system is detectable. Hence, define

$$P \triangleq \lim_{k\to\infty} P_{k|k-1}, K \triangleq PC^T(CPC^T + R)^{-1}.$$

Since control systems usually run for a long time, we can assume that the system is already in steady state. The covariance of the initial state is assumed $\Sigma = P$. Hence, the Kalman filter can be rewritten as follows:

$$\begin{aligned}
&\hat{x}_{0|-1} = \bar{x}_0, \hat{x}_{k+1|k} = A\hat{x}_{k|k} + Bu_k,\\
&\hat{x}_{k|k} = \hat{x}_{k|k-1} + K(y_k - C\hat{x}_{k|k-1}).
\end{aligned}$$

Based on the optimal state estimate $\hat{x}_k$, the optimal control input u_k^* is provided by the LQG controller:

$$u_k^* = -(B^T SB + V)^{-1} B^T SA\hat{x}_{k|k},$$

where S satisfies the following Riccati equation

$$S = A^T SA + W - A^T SB(B^T SB + V)^{-1} B^T SA.$$

Define $L \triangleq -(B^T SB + V)^{-1} B^T SA$, then

$$u_k^* = L\hat{x}_{k|k}. \tag{4}$$

The objective function given by the optimal estimator and controller in our case is

$$J = \mathrm{tr}(SQ) + \mathrm{tr}\left[(A^T SA + W - S)(P - KCP)\right]. \tag{5}$$

The χ^2 detector [23] is widely used to detect anomalies in control systems. It takes the following form at time k:

$$g_k = \sum_{i=k-\mathscr{T}+1}^{k} (y_i - C\hat{x}_{i|i-1})^T \mathscr{P}^{-1} (y_i - C\hat{x}_{i|i-1}) \underset{\mathscr{H}_1}{\overset{\mathscr{H}_0}{\lessgtr}} \eta, \tag{6}$$

where $\mathscr{T}$ is the window size of detection, $\mathscr{P} = (CPC^T + R)$ and η is the threshold which is related with the false alarm rate. When the system is under normal operation, the left of the above equation is χ^2 distributed with $m\mathscr{T}$ degrees of freedom. Furthermore, $\mathscr{H}_0$ denotes the system is under normal operation while $\mathscr{H}_1$ denotes a triggered alarm.

Here, define the probability of false alarm α_k and the probability of detection rate β_k as:

$$\alpha_k \triangleq \mathbb{P}(g_k > \eta | \mathscr{H}_0), \qquad \beta_k \triangleq \mathbb{P}(g_k > \eta | \mathscr{H}_1).$$

2.2 Attack Model

In this section, we introduce a replay attack model and analyze the feasibility of this kind of attacks on the control system.

The adversary is assumed to have the following capabilities and resources:

1. The attacker has access to all the real time sensor measurements. In other words, it knows true $y_0, \ldots, y_k$ at time k.
2. The attacker can modify the true sensory data y_k to arbitrary signals y'_k by adding malicious data y_k^a to the sensor measurement.
3. The attacker can inject attack u_k^a to the control input.

Under the above attack, the system dynamics changes to the following form:

$$x_{k+1} = Ax_k + Bu_k + B^a u_k^a + w_k, \qquad y'_k = Cx_k + D^a y_k^a + v_k. \tag{7}$$

where u_k^a and d_k^a are the harmful input and output.

Given these capabilities, the adversary can launch multiple types of attacks, such as zero dynamics attack[24], covert attack[25], false data injection attack[26, 27] and replay attack[5, 7, 8]. In this chapter, we mainly focus on replay attacks. Without loss of generality, we assume that attack starts at time 0. During the replay attack, the following attack strategies are employed:

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 0 to the recorded signals, i.e.,

$$D^a y_k^a = y_k' - Cx_k - v_k = y_{k-\Delta k} - Cx_k - v_k, \forall\, 0 \leq k \leq T_p,$$

   where $\Delta k = -k_1$.
3. The attacker inject the malicious input $B^a u_k^a$.

Here, considering the above system, detector and the attack strategies, the stability of $\mathscr{A} \triangleq (A+BL)(I-KC)$ implies that the detection rate β_k converges to the false alarm rate α_k. If $\mathscr{A}$ is unstable, the detection rate β_k goes to one. For a more detailed discussion on the detectability of replay attack, please refer to [7].

Since the classical *passive* detection scheme, where the detector passively observes the sensory data, is incapable of detection a replay attack in some CPS, an *active* detection scheme is needed to solve the problem. In the following section, we will develop a physical watermark scheme by which the detector can better detect such attacks.

3 Physical Watermark Scheme

The main idea of physical watermark is to inject a random noise, 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 order to detect replay attack, the controller is redesigned as

$$u_k = u_k^* + \Delta u_k, \tag{8}$$

where u_k^* is the optimal LQG control signal and Δu_k is drawn from an IID Gaussian distribution with zero mean and covariance $\mathscr{Q}$, and the watermark signal sequence are chosen to be also independent of u_k^*.

3.1 LQG Performance Loss

Δu_k is added as an authentication signal. It is chosen to be zero mean because we do not wish to introduce any bias to x_k. It is clear that when there is no attack, the controller is not optimal in the LQG sense anymore, which means that in order to detect the attack, we need to sacrifice control performance. The following theorem characterizes the loss of LQG performance when we inject Δu_k into the system.

Theorem 1 ([7]). *The LQG performance after adding Δu_k is given by*

$$J' = J + \underbrace{\mathrm{tr}[(V + B^T S B)\mathscr{Q}]}_{\Delta J}. \tag{9}$$

3.2 Detection Performance

Consider the χ^2 detector after adding the watermarking signal. The following theorem shows the effectiveness of the detector under the modified control scheme.

Theorem 2 ([7]). *In the absence of an attack,*

$$\mathbb{E}[(y_k - Cx_{k|k-1})^T \mathscr{P}^{-1}(y_k - Cx_{k|k-1})] = m. \tag{10}$$

Under attack

$$\lim_{k\to\infty} \mathbb{E}[(y'_k - Cx_{k|k-1})^T \mathscr{P}^{-1}(y_k - Cx_{k|k-1})] = m + 2\,\mathrm{tr}(C^T \mathscr{P}^{-1} C \mathscr{U}),$$

where $\mathscr{U}$ is the solution of the following Lyapunov equation

$$\mathscr{U} - B\mathscr{Q}B^T = \mathscr{A}\mathscr{U}\mathscr{A}^T.$$

Corollary 1 ([7]). *In the absence of an attack,*

$$\mathbb{E}[(y_k - Cx_{k|k-1})^T \mathscr{P}^{-1}(y_k - Cx_{k|k-1})] = m\mathscr{T}. \tag{11}$$

Under attack

$$\lim_{k\to\infty} \mathbb{E}[(y'_k - Cx_{k|k-1})^T \mathscr{P}^{-1}(y_k - Cx_{k|k-1})] = m\mathscr{T} + 2\,\mathrm{tr}(C^T \mathscr{P}^{-1} C \mathscr{U})\,\mathscr{T}.$$

3.3 The Trade-off between Control and Detection Performance

The authentication signal Δu_k can be optimized such to maximize the detection performance while minimizing the effect on controller performance. As the authen-

tication signal has to be zero mean, the design hinges on the covariance matrix $\mathscr{Q}$. Let the optimal value of $\mathscr{Q}$, based on the design requirements, be denoted by $\mathscr{Q}^*$.

The optimization problem can be setup in two ways. Initially, the LQG performance loss (ΔJ) can be constrained to be less than some design parameters Θ, and the increase (Δg_k) in the expected value of the quadratic residues in case of an attack maximized. In this case, the optimal $\mathscr{Q}^*$ is the solution to the following optimization problem

$$\begin{aligned} \arg\max_{\mathscr{Q}} \quad & \operatorname{tr}(C^T \mathscr{P}^{-1} C \mathscr{U}) \\ \text{subject to} \quad & \mathscr{U} - B\mathscr{Q}B^T = \mathscr{A}\mathscr{U}\mathscr{A}^T \\ & \mathscr{Q} \geq 0 \\ & \operatorname{tr}[(V + B^T S B)\mathscr{Q}] \leq \Theta. \end{aligned} \tag{12}$$

Remark 1. It can be observed from Theorem 1 and Theorem 2 that the increase (ΔJ) in LQG cost and increase (Δg_k) in the expectation of the quadratic residues are linear functions of the noise covariance matrix matrix $\mathscr{Q}$. Thus, the optimization problem is a semidefinite programming problem, and hence can be solved efficiently.

Theorem 3 ([8]). *There exists an optimal $\mathscr{Q}^*$ for* (12) *of the following form:*

$$\mathscr{Q}^* = \alpha \omega \omega^T,$$

where $\alpha > 0$ is a scalar and ω is a vector such that $\omega^T \omega = 1$.

Another way of optimizing is to constrain the increase (Δg_k) in the expected values of the quadratic residues to be above a fixed value Γ, thereby guaranteeing a certain rate of detection, and the performance loss (ΔJ) can be minimized. The optimal Q is now the solution to the optimization problem

$$\begin{aligned} \arg\max_{\mathscr{Q}} \quad & \operatorname{tr}[(V + B^T S B)\mathscr{Q}] \\ \text{subject to} \quad & \mathscr{U} - B\mathscr{Q}B^T = \mathscr{A}\mathscr{U}\mathscr{A}^T \\ & \mathscr{Q} \geq 0 \\ & \operatorname{tr}(C^T \mathscr{P}^{-1} C \mathscr{U}) \geq \Gamma. \end{aligned} \tag{13}$$

Remark 2. The solutions of the two optimization problems given in (12) and (13) will be scalar multiples of each other, thus solving either optimization problem guarantees same performance.

4 Extensions of Physical Watermark Scheme

4.1 A Non-IID Watermarking Design Approach

In this subsection, we further investigate the problem of designing the watermarking signal to achieve the optimal trade-off between control performance and detection performance. The following technique generalizes the results in [7, 8] and considers non-independent and identically distributed Gaussian process.

For the sake of simplicity, we define $\zeta_k \triangleq \Delta u_k$, where Δu_k is defined in (8). Correspondingly, (8) is rewritten as:

$$u_k = u_k^* + \zeta_k. \tag{14}$$

Here, the auto-covariance function is defined as:

$$\Gamma(d) \triangleq \mathrm{Cov}(\zeta_0, \zeta_k) = \mathbb{E}\zeta_0 \zeta_d^T,$$

and the watermarking signal is generated by a Hidden-Markov Model (HMM)

$$\xi_{k+1} = A_h \xi_k + \varphi_k, \qquad \zeta_k = C_h \xi_k, \tag{15}$$

where $\varphi_k \in \mathbb{R}^{n_h}$, $k \in \mathbb{Z}$ is a sequence of IID zero-mean Gaussian random variables with covariance Ψ, and $\xi_k \in \mathbb{R}^{n_h}$ is the hidden state. To make ζ_k be a stationary process, the covariance of ξ_0 is assumed to be the solution of the following Lyapunov equation

$$\mathrm{Cov}(\xi_0) = A_h \mathrm{Cov}(\xi_0) A_h^T + \Psi,$$

where A_h is strictly stable. It is assumed that the watermark signal is chosen from a HMM with $\rho(A_h) < \rho$, where $\rho < 1$ is a design parameter. A value of ρ close to 1 gives the system operator more freedom to design the watermark signal, while a value of ρ close to 0 improves the freshness of the watermark signal by reducing the correlation of φ_k at different time steps. To simplify notations, define the feasible set $\mathscr{G}(\rho)$ as

$$\mathscr{G}(\rho) = \{\Gamma : \Gamma \text{ is generated by an HMM (15) with } \rho(A_h) < \rho\}.$$

4.1.1 LQG Performance

Similarly, the injection of the watermarking signal ζ_k degrades the LQG performance. The LQG cost is $J = \lim E \frac{1}{2T+1} \left[\sum_{k=-T}^{T} (x_k^T W x_k + u_k^T V u_k)\right]$. The following theorem characterizes the performance loss incurred by the additional watermark.

Theorem 4 ([5]). *The LQG performance of the system described by* (1) (2) *and* (14) *is characterized as:*

$$J = J^* + \Delta J,$$

where J^* is the optimal LQG cost without the watermark signal and

$$\Delta J = \mathrm{tr}\left\{ V\Gamma(0) + 2V\,\mathrm{sym}\left[L\sum_{d=0}^{\infty}(A+BL)^d B\Gamma(1+d)\right]\right\} + \mathrm{tr}\left[(W+L^TVL)\Theta_1\right], \tag{16}$$

where

$$\Theta_1 \triangleq 2\sum_{d=0}^{\infty}\mathrm{sym}\left[(A+BL)^d\mathscr{L}_1(\Gamma(d))\right] - \mathscr{L}_1(\Gamma(0)),$$

and $\mathscr{L}_1 : \mathbb{C}^{p\times p} \to \mathbb{C}^{n\times n}$ is a linear operator defined as

$$\mathscr{L}_1(X) = \sum_{i=0}^{\infty}(A+BL)^i BXB^T((A+BL)^i)^T = (A+BL)\mathscr{L}_1(X)(A+BL)^T + BXB^T.$$

4.1.2 Detection Performance

In the absence of the attack, since the real time authentication signal ζ_k and the residue z_k are available to the detector of the system, the residue z_k follows a Gaussian distribution with mean zero and covariance $\mathscr{P} = CPC^T + R$ [23].

In the presence of the attack, the residue z_k converges to a Gaussian with mean μ_{k-1} and covariance $(\mathscr{P}+\Sigma)$ [5], where

$$\mu_k \triangleq -C\sum_{i=0}^{k}\mathscr{A}^{k-i}B\zeta_i \text{ and } \Sigma = 2\sum_{d=0}^{\infty}C\,\mathrm{sym}[\mathscr{A}^d\mathscr{L}_2(\Gamma(d))]C^T - \mathscr{L}_2(\Gamma(0))]C^T, \tag{17}$$

where $\mathscr{L}_2 : \mathbb{C}^{p\times p} \to \mathbb{C}^{n\times n}$ is a linear operator on the space of $p\times p$, which is defined as

$$\mathscr{L}_2(X) \triangleq \sum_{i=0}^{\infty}\mathscr{A}^i BXB^T(\mathscr{A}^i)^T = \mathscr{A}^i\mathscr{L}_2(X)(\mathscr{A}^i)^T + BXB^T.$$

To detect the replay attack, we need a detector to differentiate the distribution of y_k under the following two hypotheses:

$$\mathscr{N}_0 : z_k \sim \mathscr{N}_0(0,\mathscr{P}); \qquad \mathscr{N}_1 : z_k \sim \mathscr{N}_1(\mu_{k-1},\mathscr{P}+\Sigma).$$

By the Neyman-Pearson lemma [28], the optimal detector is given by the Neyman-Pearson detector as discussed in the following theorem.

Theorem 5 ([5]). *The optimal Neyman-Pearson detector rejects $\mathcal{N}_0$ in favor of $\mathcal{N}_1$ if*

$$g_{NP}(z_k, \zeta_{k-1}, \zeta_{k-2}, \cdots) = z_k^T \mathscr{P}^{-1} z_k - (z_k - \mu_{k-1})^T (\mathscr{P} + \Sigma)^{-1} (z_k - \mu_{k-1}) \geq \eta. \tag{18}$$

Otherwise, hypothesis $\mathscr{H}_0$ is accepted.

Since the detection rate and expected time to detection involve integrating a Gaussian distribution, which usually does not have an analytical solution, the Kullback-Leibler (KL) divergence is used to characterize the detection performance. The following theorem quantifies the detection performance from the perspective of the expected KL divergence between $\mathcal{N}_0$ and $\mathcal{N}_1$:

Theorem 6 ([5]). *The expected KL divergence of distribution $\mathcal{N}_1$ and $\mathcal{N}_0$ is*

$$\mathbb{E}\, D_{KL}(\mathcal{N}_1 \| \mathcal{N}_0) = \operatorname{tr}\left(\Sigma \mathscr{P}^{-1}\right) - \frac{1}{2} \log\det\left(I + \Sigma \mathscr{P}^{-1}\right). \tag{19}$$

Furthermore, the expected KL divergence satisfies the inequality

$$\frac{1}{2} \operatorname{tr}\left(\Sigma \mathscr{P}^{-1}\right) \leq \mathbb{E}\, D_{KL}(\mathcal{N}_1 \| \mathcal{N}_0) \leq \operatorname{tr}\left(\Sigma \mathscr{P}^{-1}\right) - \frac{1}{2} \log\left[1 + \operatorname{tr}\left(\Sigma \mathscr{P}^{-1}\right)\right]. \tag{20}$$

where the upper bound is tight if C is of rank 1.

4.1.3 The Optimal Watermarking Signal

In order to achieve the optimal tradeoff between the control performance and detection performance, the optimization problem is formulated as follows:

$$\begin{aligned} \underset{\Gamma(d) \in \mathscr{G}(\rho)}{\arg\max} \quad & \mathbb{E}\, D_{KL}(\mathcal{N}_1 \| \mathcal{N}_0), \\ \text{subject to} \quad & \Delta J \leq \delta. \end{aligned} \tag{21}$$

where $\delta > 0$ is a design parameter depending on how much control performance loss is tolerable.

It is worth noticing that directly maximizing the detection performance is computationally difficult. Notice that the expected KL divergence is relaxed to $\operatorname{tr}\left(\Sigma \mathscr{P}^{-1}\right)$, using the upper and lower bound derived in Theorem 6. One can transform the above optimization problem to following problem:

$$\begin{aligned} \underset{\Gamma(d) \in \mathscr{G}(\rho)}{\arg\max} \quad & \operatorname{tr}\left(\Sigma \mathscr{P}^{-1}\right), \\ \text{subject to} \quad & \Delta J \leq \delta. \end{aligned} \tag{22}$$

Although Σ and J are linear functions of Γ, convex optimization techniques cannot be directly applied to solve (22), since Γ is in an infinite dimensional space.

Therefore, (22) is transformed into the frequency domain. Before continuing on, the following definition is needed.

Definition 1 ([5]). ν is a positive Hermitian measure of size $p \times p$ on the interval $(-0.5, 0.5]$ if for a Borel set $S_B \subseteq (-0.5, 0.5]$, $\nu(S_B)$ is a positive semidefinite Hermitian matrix with size $p \times p$.

The following theorem establishes the existence of a frequency domain representation for $\Gamma(d)$.

Theorem 7 (Bochner's Theorem [29, 30]). *$\Gamma(d)$ is the autocovariance function of a stationary Gaussian process ζ_k if and only if there exists a unique positive Hermitian measure ν of size $p \times p$, such that*

$$\Gamma(d) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \exp(2\pi j d\omega) d\nu(\omega). \tag{23}$$

By the fact that $\Gamma(d)$ is real, the Hermitian measure ν satisfies the following property, which can be applied to the Fourier transform of the real valued signals.

Proposition 1 ([5]). *$\Gamma(d)$ is real if and only if for all Borel-measureable sets $S_B \subseteq (-0.5, 0.5]$,*

$$\nu(S_B) = \overline{\nu(-S_B)}. \tag{24}$$

By (24), (23) can be simplified as

$$\Gamma(d) = 2\Re\left(\int_{0}^{\frac{1}{2}} \exp(2\pi j d\omega) d\nu(\omega) \right).$$

Theorem 8 ([5]). *The optimal solution (not necessarily unique) of* (22) *is*

$$\Gamma_*(d) = 2\rho^{|d|} \Re\left(\exp(2\pi j d\omega_*) H_*\right), \tag{25}$$

where ω_* and H_* are the solution of the ensuing optimization problem.

$$\begin{aligned} \arg\max_{\omega, H} \quad & \operatorname{tr}[\mathscr{F}_2(\omega, H) C^T \mathscr{P}^{-1} C], \\ \textit{subject to} \quad & \mathscr{F}_1(\omega, H) \leq \delta, 0 \leq \omega \leq 0.5, \\ & H \textit{ Hermitian and Positive Semidefinite}, \end{aligned} \tag{26}$$

where the function $\mathscr{F}_1$ and $\mathscr{F}_2$ are defined as

$$\mathscr{F}_1(\omega, H) \triangleq tr[V\Theta_2] + \operatorname{tr}[(W + L^T V L)\Theta_3], \tag{27}$$

$$\mathscr{F}_2(\omega, H) \triangleq 2\Re\{2\,\mathrm{sym}[(I - s\rho\mathscr{A})^{-1}\mathscr{L}_2(H)] - \mathscr{L}_2(H)\}, \tag{28}$$

where

$$\Theta_2 \triangleq 2\Re\{2\,\mathrm{sym}(s\rho L[I - s\rho(A+BL)]^{-1}BH) + H\},$$
$$\Theta_3 \triangleq 2\Re\{2\,\mathrm{sym}[(I - s\rho(A+BL))^{-1}\mathscr{L}_1(H)] - \mathscr{L}_1(H)\},$$
$$s \triangleq \exp(2\pi j\omega).$$

Furthermore, one optimal H_* of optimization problem (26) is of the form $H_* = hh^H$, where $h \in \mathbb{C}^p$. The corresponding HMM is given by

$$\xi_{k+1} = \rho \begin{bmatrix} \cos 2\pi\omega_* & -\sin 2\pi\omega_* \\ \sin 2\pi\omega_* & \cos 2\pi\omega_* \end{bmatrix} \xi_k + \psi_k, \qquad \zeta_k = \begin{bmatrix} \sqrt{2}h_r & \sqrt{2}h_i \end{bmatrix} \xi_k, \tag{29}$$

where $h_r, h_i \in \mathbb{R}^p$ are the real and imaginary part of h respectively and $\Psi = \mathrm{Cov}(\psi_k) = (1-\rho^2)I$.

4.2 An On-line Design Approach

It is worth noticing that in order to design 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 is our focus in this section. Based on the physical watermark 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 marked parameters in Fig. 1: the covariance U_k of the watermark signal ϕ_k and the optimal detector based on the estimated parameters.

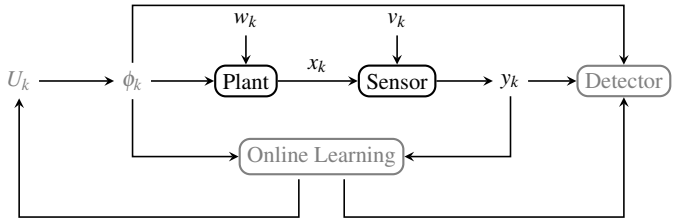

Fig. 1 The system diagram.

To simplify notations, in this subsection we consider a stable open-loop system. The LTI system described by (1) (2) is rewritten as follows:

$$x_k = Ax_{k-1} + B\phi_k + w_k, \tag{30}$$
$$y_k = Cx_k + v_k, \tag{31}$$

where $\phi_k \in \mathbb{R}^p$ is the watermark signal and its covariance is denoted as U.

4.2.1 Physical Watermark for Systems with Known Parameters

The analyses on the control performance, detection performance and optimal problem are similar to that in the above section. Due to the space constraints, we only provide the outline. Please refer to [15] for more details.

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

$$y_k = \varphi_k + \vartheta_k, \tag{32}$$

where

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

where $H_\tau = CA^\tau B$. It is easy to know that φ_k is a zero mean Gaussian whose covariance converges to $\mathscr{U}$, where $\mathscr{U} \triangleq \sum_{\tau=0}^{\infty} H_\tau U H_\tau^T$. Similarly, ϑ_k is a zero mean Gaussian noise whose covariance is $\mathscr{W} = CPC^T + R$.

Under the replay attack, the replayed y'_k can be written as

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

Since Δk is unknown to the system operator, we shall treat $\varphi_{k-\Delta k}$ as a zero mean Gaussian random variable with covariance $\mathscr{U}$. As a result, y'_k is a zero mean Gaussian random variable with covariance $\mathscr{U} + \mathscr{W}$.

Here, we provide the following two hypotheses on the distribution of y_k:

$$\mathscr{H}_0 \colon y_k \sim \mathscr{N}_0(\varphi_k, \mathscr{W}), \qquad \mathscr{H}_1 \colon y_k \sim \mathscr{N}_1(0, \mathscr{U} + \mathscr{W}).$$

The Neyman-Pearson detector [28] is employed to differentiate two distributions and the KL divergence is used to characterize the detection performance. Hence, we aim to maximize $\mathrm{tr}(\mathscr{U}\mathscr{W}^{-1})$ to maximize the detection performance.

Correspondingly, the following LQG metric is used to quantify the performance loss:

$$J = \lim_{T \to +\infty} \mathbb{E}\left(\frac{1}{T} \sum_{k=0}^{T-1} \begin{bmatrix} y_k \\ \phi_k \end{bmatrix}^T X \begin{bmatrix} y_k \\ \phi_k \end{bmatrix} \right) = \mathrm{tr}(X_{yy}\mathscr{W}) + \mathrm{tr}(XS), \tag{33}$$

where

$$S = \begin{bmatrix} \mathscr{U} & H_0 U \\ U H_0^T & U \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} X_{yy} & X_{y\phi} \\ X_{\phi y} & X_{\phi\phi} \end{bmatrix} > 0$$

is the weight matrix for the LQG control, which is chosen by the system operator.

Therefore, in order the achieve the optimal trade-off between the control and detection performance, the optimization problem is formulated as follows:

$$\begin{aligned} U_* = \arg\max_{U\geq 0} \quad & \operatorname{tr}(\mathscr{U}\mathscr{W}^{-1}) \\ \text{subject to} \quad & \operatorname{tr}(XS) \leq \delta, \end{aligned} \tag{34}$$

where δ is a design parameter.

An important property of the optimization problem (34) is that the optimal solution is usually a rank-1 matrix, which is formalized by the following theorem:

Theorem 9 ([15]). *The optimization problem* (34) *is equivalent to*

$$\begin{aligned} U_* = \arg\max_{U\geq 0} \quad & \operatorname{tr}(U\mathscr{P}) \\ \textit{subject to} \quad & \operatorname{tr}(U\mathscr{X}) \leq \delta, \end{aligned} \tag{35}$$

where

$$\mathscr{P} = \sum_{\tau=0}^{\infty} H_\tau^T \mathscr{W}^{-1} H_\tau \quad \textit{and} \quad \mathscr{X} = \left(\sum_{\tau=0}^{\infty} H_\tau^T X_{yy} H_\tau\right) + H_0^T X_{y\phi} + X_{\phi y} H_0 + X_{\phi\phi}.$$

The optimal solution to (35) *is $U_* = zz^T$, where z is the eigenvector corresponding to the maximum eigenvalue of the matrix $\mathscr{X}^{-1}\mathscr{P}$ and $z^T\mathscr{X}z = \delta$. Furthermore, the solution is unique if $\mathscr{X}^{-1}\mathscr{P}$ has only one maximum eigenvalue.*

Then we will develop 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 this subsection, we make the following assumptions:

Assumption 1 *[15]*

1. *A is diagonalizable.*
2. *The maximum eigenvalue of $\mathscr{X}^{-1}\mathscr{P}$ 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.*

4.2.2 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. A pseudo-code form for Algorithm 1 is as follows:

Then we will introduce this algorithm in detail.

Algorithm 1 Online Watermarking Design

Initialization: $\mathscr{P}_{-1} \leftarrow I$, $\mathscr{X}_{-1} \leftarrow X_{\phi\phi}, k \leftarrow 0$
Iteration:
while true **do**
  $U_{k,*} \leftarrow \arg\max_{U\geq 0, \mathrm{tr}(U\mathscr{X}_{k-1})\leq\delta} \mathrm{tr}(U\mathscr{P}_{k-1})$
  $U_k \leftarrow U_{k,*} + (k+1)^{-\upsilon}\delta I$
  Generate random variable $\zeta_k \sim \mathscr{N}(0,I)$
  Apply watermark signal $\phi_k \leftarrow U_k^{1/2}\zeta_k$
  Collect sensory data y_k
  $H_{k,\tau} \leftarrow \frac{1}{k-\tau+1}\sum_{t=\tau}^{k} y_t \phi_{t-\tau}^T U_{t-\tau}^{-1}$
  Compute the coefficient of $p_k(x)$ by solving (40)
  if $p_k(x)$ is Schur stable **then**
    Update $\mathscr{P}_k, \mathscr{X}_k$ from (41)-(46)
  end if
  Update $\hat{g}_k$ from (47)
  $k \leftarrow k+1$
end while

Generation of the Watermark Signal ϕ_k

Let us design U_k, which can be considered as an approximation for the optimal covariance of the watermark signal U, as

$$U_k = U_{k,*} + \frac{\delta}{(k+1)^{\upsilon}} I, \tag{36}$$

where $0 < \upsilon < 1$, δ is the maximum tolerable LQG loss, and $U_{k,*}$ is the solution of the following optimization problem

$$\begin{aligned} U_{k,*} = \arg\max_{U\geq 0} \quad & \mathrm{tr}(U\mathscr{P}_{k-1}), \\ \text{subject to} \quad & \mathrm{tr}(U\mathscr{X}_{k-1}) \leq \delta, \end{aligned} \tag{37}$$

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

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

Inference on H_τ

Define the following quantity $H_{k,\tau}$, where $0 \leq \tau \leq 3\tilde{n} - 2$, as

$$H_{k,\tau} \triangleq \frac{1}{k-\tau+1}\sum_{t=\tau}^{k} y_t \phi_{t-\tau}^T U_{t-\tau}^{-1}$$
$$= H_{k-1,\tau} + \frac{1}{k-\tau+1}\left(y_k \phi_{k-\tau}^T U_{k-\tau}^{-1} - H_{k-1,\tau}\right), \tag{38}$$

where $H_{k,\tau}$ can be interpreted as an estimate of H_τ.

It is worth noticing that the calculation of the matrices $\mathscr{U}$, $\mathscr{W}$, $\mathscr{P}$ and $\mathscr{X}$ 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 1. *Assuming the matrix A is diagonalizable with $\lambda_1, \ldots, \lambda_{\tilde{n}}$ being its distinct eigenvalues, then there exist unique $\Omega_1, \cdots, \Omega_{\tilde{n}}$, such that $H_\tau = \sum_{i=1}^{\tilde{n}} \lambda_i^\tau \Omega_i$.*

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 $i \geq 0$:

$$H_{i+\tilde{n}} + \alpha_{\tilde{n}-1}H_{i+\tilde{n}-1} + \cdots + \alpha_0 H_i = CA^i p(A)B = 0. \tag{39}$$

Leveraging (39), 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:

$$\begin{bmatrix} \alpha_{k,0} \\ \vdots \\ \alpha_{k,\tilde{n}-1} \end{bmatrix} \triangleq -\Xi_k^{-1} \begin{bmatrix} \mathrm{tr}(\mathscr{H}_{k,0}^T \mathscr{H}_{k,\tilde{n}}) \\ \vdots \\ \mathrm{tr}(\mathscr{H}_{k,\tilde{n}-1}^T \mathscr{H}_{k,\tilde{n}}) \end{bmatrix}, \tag{40}$$

where

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

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

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

and we shall estimate Ω_i as

$$\begin{bmatrix} \Omega_{k,1} \\ \vdots \\ \Omega_{k,\tilde{n}} \end{bmatrix} = (V_k \otimes I_m)^+ \begin{bmatrix} H_{k,0} \\ \cdots \\ H_{k,3\tilde{n}-2} \end{bmatrix}. \tag{41}$$

Inference on φ_k, ϑ_k and $\mathscr{W}$

Define

$$\hat{\varphi}_k \triangleq \sum_{i=1}^{\tilde{n}} \hat{\varphi}_{k,i}, \tag{42}$$

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

$$\hat{\vartheta}_k \triangleq y_k - \hat{\varphi}_k. \tag{43}$$

The covariance of ϑ_k can be estimated as

$$\mathscr{W}_k \triangleq \frac{1}{k+1}\sum_{t=0}^{k} \hat{\vartheta}_t \hat{\vartheta}_t^T. \tag{44}$$

Inference on $\mathscr{P}$, $\mathscr{X}$, $\mathscr{U}$ and g_k

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

$$\begin{aligned}\mathscr{P}_k &= \sum_{\tau=0}^{\infty}\left(\sum_{i=1}^{\tilde{n}}\lambda_{k,i}^{\tau}\Omega_{k,i}\right)^T \mathscr{W}_k^{-1}\left(\sum_{i=1}^{\tilde{n}}\lambda_{k,i}^{\tau}\Omega_{k,i}\right)\\ &= \sum_{i=1}^{\tilde{n}}\sum_{j=1}^{\tilde{n}}\frac{1}{1-\lambda_{k,i}\lambda_{k,j}}\Omega_{k,i}^T\mathscr{W}_k^{-1}\Omega_{k,j},\end{aligned} \tag{45}$$

and

$$\begin{aligned}\mathscr{X}_k &= \sum_{\tau=0}^{\infty}\left(\sum_{i=1}^{\tilde{n}}\lambda_{k,i}^{\tau}\Omega_{k,i}\right)^T X_{yy}\left(\sum_{i=1}^{\tilde{n}}\lambda_{k,i}^{\tau}\Omega_{k,i}\right) + \sum_{i=1}^{\tilde{n}}\Omega_{k,i}^T X_{y\phi} + X_{\phi y}\sum_{i=1}^{\tilde{n}}\Omega_{k,i} + X_{\phi\phi}\\ &= \sum_{i=1}^{\tilde{n}}\sum_{j=1}^{\tilde{n}}\frac{1}{1-\lambda_{k,i}\lambda_{k,j}}\Omega_{k,i}^T X_{yy}\Omega_{k,j} + \sum_{i=1}^{\tilde{n}}\Omega_{k,i}^T X_{y\phi} + X_{\phi y}\sum_{i=1}^{\tilde{n}}\Omega_{k,i} + X_{\phi\phi}.\end{aligned} \tag{46}$$

The Neyman-Pearson detection statistics g_k can be approximated by

$$\hat{g}_k = (y_k - \hat{\varphi}_k)^T \mathscr{W}_k^{-1}(y_k - \hat{\varphi}_k) - y_k^T(\mathscr{W}_k + \mathscr{U}_k)^{-1} y_k, \tag{47}$$

where

$$\mathscr{U}_k = \sum_{\tau=0}^{\infty} \left(\sum_{i=1}^{\tilde{n}} \lambda_{k,i}^{\tau} \Omega_{k,i} \right) U_{k,*} \left(\sum_{i=1}^{\tilde{n}} \lambda_{k,i}^{\tau} \Omega_{k,i} \right)^T$$
$$= \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{\tilde{n}} \frac{1}{1-\lambda_{k,i}\lambda_{k,j}} \Omega_{k,i} U_{k,*} \Omega_{k,j}^T. \tag{48}$$

4.2.3 Algorithm Properties

The following theorem establishes the convergence of $U_{k,*}$ and g_k, the proof can be found in [15].

Theorem 10. *Assuming that A is strictly stable and Assumption 2 holds. If* $0 < \upsilon < 1$*, then for any* $\varepsilon > 0$*, the following limits hold almost surely:*

$$\lim_{k\to\infty} \frac{U_{k,*} - U_*}{k^{-\gamma+\varepsilon}} = 0, \lim_{k\to\infty} \frac{\hat{g}_k - g_k}{k^{-\gamma+\varepsilon}} = 0, \tag{49}$$

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

4.2.4 Simulation Result

In this section, the performance of the proposed algorithm is evaluated. We will apply the proposed online "learning" approach to a numerical example. First we choose $m = 3, n = 5, p = 2$ and A, B, C are all randomly generated, with A being stable. It is assumed that X in (33), the covariance matrices Q and R are all identity matrices with proper dimensions. We assume that δ in (35) is equal to 10% of optimal LQG cost J_0. Fig. 2 shows relative error $\|U_{k,*} - U_*\|_F / \|U_*\|_F$ of the estimated $U_{k,*}$ v.s. time k for different υs.

From Fig 2, one can see that the estimator error converges to 0 as time k goes to infinity and the convergence approximately follows a power law. From Theorem 10, we know that $U_{k,*} - U_* \sim O(k^{-\gamma+\varepsilon})$, where $\gamma = (1-\upsilon)/2$. However, from Fig 2, it seems that the convergence speed of the error for different υ is comparable. Notice that Theorem 10 only provides an upper bound for the convergence rate. As a result, it would be interesting to quantify the exact impact of υ on the convergence rate, which we shall leave as a future research direction.

Now we consider the detection performance of our online watermark signal design, after an initial inference period, where no attack is present. It is assumed that the attacker records the sensor readings from time $10^4 + 1$ to $10^4 + 100$ and replays them to the system from time $10^4 + 101$ to $10^4 + 200$. Fig 3 shows the trajectory of the Neyman-Pearson statistic g_k and our estimate $\hat{g}_k$ of g_k for one simulation. Notice that $\hat{g}_k$ can track g_k with high accuracy. Furthermore, both $\hat{g}_k$ and g_k are significantly larger when the system is under replay attack (after time $10^4 + 101$). Hence one can

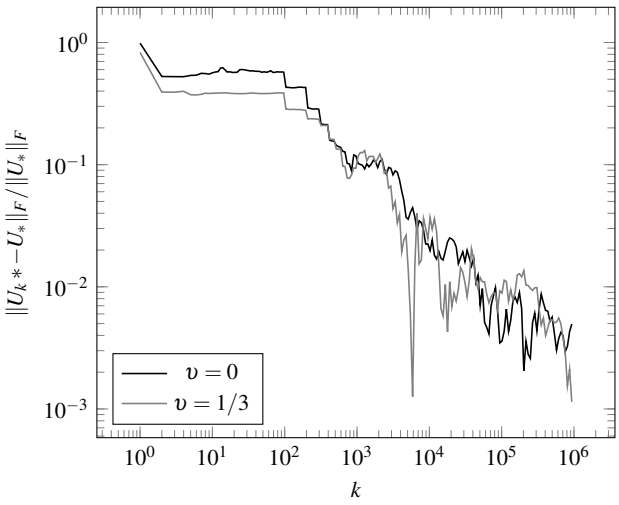

Fig. 2 Relative error of $U_{k,*}$ for different υ. The black solid line denotes the relative error of $U_{k,*}$ when $\upsilon = 0$. The gray solid line is the relative error of $U_{k,*}$ when $\upsilon = 1/3$.

conclude that even without parameter knowledge, we can successfully estimate g_k and detect the presence of the replay attack.

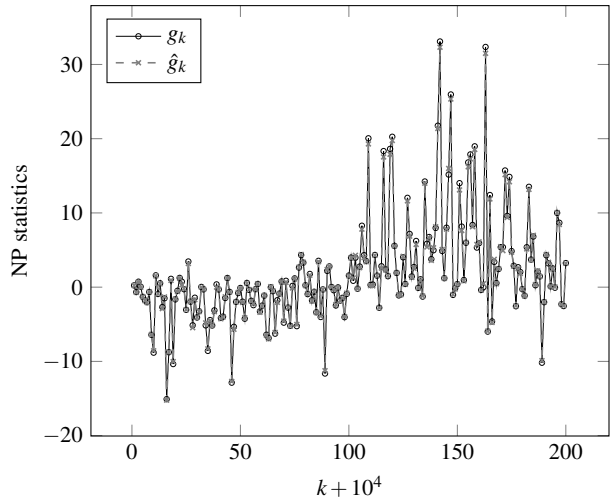

Fig. 3 The detection statistics v.s. time. The black solid line with circle markers is the true Neyman-Pearson statistics g_k, assuming full system knowledge. The gray dashed line with cross markers denotes our estimated $\hat{g}_k$.

4.3 A Multiplicative Watermarking Design

Different from the additive physical watermark scheme, where the watermarking signal is injected to the control input in the above work, Riccardo M.G. Ferrari and André M.H. Teixeira proposed a multiplicative sensor watermarking scheme. It has been applied to detect several types of attacks including replay attacks [20], routing attacks [21] and false data injection attacks [22].

In this subsection, we mainly introduce the multiplicative watermarking scheme proposed in [20], in which each sensors output is separately watermarked. Correspondingly, the equalizing filters are equipped to reconstruct the real output signal from the watermarked data. About the proofs of theorems in this subsection, please refer to [20].

Consider the following system model is as follows:

$$\begin{aligned}
\mathscr{P}&:\begin{cases} x_p(k+1) = A_p x_p(k) + B_p u(k) + \eta(k) \\ y_p(k) = C_p x_p(k) + \xi(k) \end{cases} \\
\mathscr{C}&:\begin{cases} x_{cr}(k+1) = A_c x_c(k) + B_c \tilde{y}_p(k) \\ u(k) = C_c x_c(k) + D_c \tilde{y}_p(k) \end{cases} \\
\mathscr{R}&:\begin{cases} x_r(k+1) = A_r x_r(k) + B_r u(k) + K_r \tilde{y}_p(k) \\ y_r(k) = C_r x_r(k) + D_r u(k) + E_r \tilde{y}_p(k) \end{cases},
\end{aligned} \tag{50}$$

where all notations' meaning and relative assumptions could be found in [20] and we omit them due to the space constraints. Here, Define $x_{c,r}(k) = [x_c(k)^T x_r(k)^T]^T$, the controller and detector dynamics can be represented as

$$\mathscr{F}_{cr}:\begin{cases} x_{cr}(k+1) = A_{cr} x_{cr}(k) + B_{cr} \tilde{y}_p(k) \\ y_r(k) = C_{cr} x_{cr}(k) + D_{cr} \tilde{y}_p(k) \\ u(k) = C_u x_{cr}(k) + D_u \tilde{y}_p(k) \end{cases}. \tag{51}$$

4.3.1 Multiplicative Watermarking and Equalizing Scheme

The main idea of a multiplicative sensor watermarking scheme is to pre-process the measurements through a filter parameterized by θ before transmitting them, denoted as sensor watermarking, and then to pre-process the received watermarked data through an equalizer filter parameterized by the very same θ before feeding them to the controller and anomaly detector, denoted as equalization [22]. Here, $\theta(k)$ is designed as a piecewise constant variable $\theta(k) \triangleq \theta_j \in \Theta$, for $k_j \leq k < k_{j+1}$, where $\mathscr{K}_\theta \triangleq \{k_1, \cdots, k_j, \cdots\}$ denotes the set of switching times and $\Theta \triangleq \{\theta_1, \cdots, \theta_M\}$ is the set of possible parameters [20].

For the watermarking step, the corresponding filters are denoted as $\mathscr{W}(\theta)$ and the watermarked measurements are denoted as $y_{pw}(k)$. For the equalization step, the equalizing filters are denoted as $\mathscr{Q}(\theta)$:

$$\begin{aligned}\mathscr{W}:&\begin{cases}x_w(k+1)=A_w(\theta)x_w(k)+B_w(\theta)y_p(k)\\ y_{pw}(k)=C_w(\theta)x_w(k)+D_w(\theta)y_p(k)\end{cases},\\ \mathscr{Q}:&\begin{cases}x_q(k+1)=A_q(\theta)x_q(k)+B_q(\theta)\tilde{y}_{pw}(k)\\ y_{pq}(k)=C_q(\theta)x_q(k)+D_q(\theta)\tilde{y}_{pw}(k)\end{cases},\end{aligned} \tag{52}$$

where $y_{pw}(k)$ and $\tilde{y}_{pw}(k)$ are employed to differentiate the watermarked data and the data received by the controller and anomaly detector.

Then we will introduce how to design the parameters in this scheme. For the sake of simplicity and without loss of generality, we suppose that there is only one sensor. The watermark generator is represented as follows:

$$y_{pw}(k)=\sum_{n=1}^{N} w_{A,(n)}y_{pw}(k-n)+\sum_{n=0}^{N} w_{B,(n)}y_p(k-n), \tag{53}$$

where $w_A=[w_{A,(1)},\cdots,w_{A,(N)}]^T\in\mathbb{R}^N$ and $w_B=[w_{B,(0)},\cdots,w_{B,(N)}]^T\in\mathbb{R}^{N+1}$ are the filter parameters.

Consider that the objective of equalizing filters are to reconstruct the sensor measurement $y(k)$, an intuitive approach is to derive the inverse of the respective filter, i.e.,

$$y_{pq}(k)=\frac{1}{w_{B,(0)}}\left(-\sum_{n=0}^{N} w_{B,(n)}y_{pq}(k-n)+\tilde{y}_{pw}(k)-\sum_{n=1}^{N} w_{A,(n)}\tilde{y}_{pw}(k-n)\right) \tag{54}$$

By using controllable canonical form, the corresponding parameters in (52) are designed as follows:

$$\begin{aligned}&A_w(\theta)=\begin{bmatrix}0_{N-1,1} & I_{N-1}\\ & w_A^T\end{bmatrix},\quad B_w=\begin{bmatrix}0_{N-1,1}\\ 1\end{bmatrix},\\ &C_w(\theta)=[\cdots w_{B,(n)}+w_{B,(0)}w_{A,(n)}\cdots],\quad \text{for}\quad n=1,\cdots,N,\quad D_w(\theta)=w_{B,(0)},\\ &A_q(\theta)=\begin{bmatrix}0_{N-1,1} & I_{N-1}\\ \frac{-1}{w_{B,(0)}}w_B^T\end{bmatrix},\quad B_q=\begin{bmatrix}0_{N-1,1}\\ \frac{1}{w_{B,(0)}}\end{bmatrix},\\ &C_q(\theta)=[\cdots -w_{A,(n)}-\frac{w_{B,(n)}}{w_{B,(0)}}\cdots],\quad \text{for}\quad n=1,\cdots,N,\quad D_q(\theta)=\frac{1}{w_{B,(0)}}.\end{aligned}$$

The following theorem characterizes the performance of the system with the multiplicative scheme under no replay attacks.

Theorem 11 ([20]). *Consider the closed-loop system with watermarked sensors described by* (50) *and* (52). *Assume that theta*(k) *is updated at times* $k\in\mathscr{K}_\theta$. *The performance of the closed-loop system equipped with sensor watermarking filters and equalizing filters is same as the performance of the nominal closed-loop system* (52) *if and only if the states of* $\mathscr{Q}(\theta)$ *and* $\mathscr{W}(\theta)$*are such that* $x_q(k)=x_w(k)$ *for all* $k\in\mathscr{K}_\theta$ *with no replay attacks.*

We now present the main result of this section regarding the detectability of replay attacks under the proposed watermarking scheme.

Theorem 12 ([20]). *Consider a replay attack that has recorded data from time $k_r = k_0 - T$ to $k_f = k_0 - T_f$, and let $\theta(k) = \theta'$ for $k_r \leq k \leq k_f$. Suppose the recorded data is replayed from time k_0 and let $\theta(k) = \theta$ for $k \geq k_0$. During the replay attack, y_r converges asymptotically to y'_r for y'_p if and only if $\theta = \theta'$.*

From Theorem 12, one can obtain that when $\theta \neq \theta'$, the undetectability of the replay attack is not guaranteed a priori, since it depends on the exogenous input y'_p.

4.3.2 Detection and Isolation of Replay Attacks

In this subsection, through the multiplicative watermarking scheme, an anomaly detector and a corresponding threshold will be derived. For more details about the isolation and identification of relay attacks, please refer to [20].

It is assumed that there is no replay attacks for $0 \leq k < k_0$, where k_0 is the start attack time. Furthermore, the variables x_p, x_{pw} and u remain bounded before being attacked. Here, (A_p, C_p) is assumed as a detectable pair [20].

The detector is designed in the following form [31]:

$$\begin{cases} \hat{x}_p(k+1) = A_p\hat{x}_p(k) + B_p u(k) + K(y_{pq}(k) - \hat{y}_p(k) \\ \quad \hat{y}_p(k) = C_p\hat{x}_p(k), \end{cases} \tag{55}$$

where $\hat{x}_p$ and $\hat{y}_p$ are estimates of x_p and y_p and the gain matrix K is chosen to satisfy that $A_r = A_p - KC_p$ is Schur. Set $x_r = \hat{x}_p$ and the estimation error $\varepsilon \triangleq x_p - \hat{x}_p$, under the scenario with attacks, the detection residual dynamics are as follows:

$$\begin{cases} \varepsilon(k+1) = A_r\varepsilon(k) - K\xi(k) + \eta(k) \\ \quad y_r(k) = C_p\varepsilon(k) + \xi(k), \end{cases} \tag{56}$$

and the detection threshold ith component is computed as

$$\bar{y}_r(k) \triangleq \alpha^i \left[\sum_{h=0}^{k-1} (\delta^i)^{k-1-i} (\bar{\eta}(h) + \|K\|\bar{\xi}(h)) + (\delta^i)^k \bar{x}_r(0) \right] + \bar{\xi}(k),$$

where α^i and δ^i are two constants such that $\|C_{p,(i)}(A_r)^k\| \leq \alpha^i(\delta^i)^k \leq \|C_{p,(i)}\| \cdot \|(A_r)^k\|$ with $C_{p,(i)}$ being the ith row of matrix C_p. Furthermore, $\bar{\eta}, \bar{x}_r(0)$ and $\bar{\xi}$ are upper bounds on the norms of, respectively, $\eta, x_r(0)$ and ξ [20].

Theorem 13. *[20] If there exists a time index $k_d > k_0$ and a component $i \in 1, ..., n_y$ such that during a cyber replay attack the following inequality holds*

$$\left| C_{p,(i)} \left[\sum_{h=k_0}^{k_d-1} (A_r)^{k_d-1-h} (B_p \Delta u(h) - K \Delta y_p(h)) \right] + \Delta y_p(k) \right|$$

$$> 2\alpha^i \sum_{h=0}^{k_d-1-h} (\delta^i)^{k_d-1-h} (\bar{\eta}(h) + \|K\| \bar{\xi}(h)) + (\delta^i)^{k_d-k_0} (\alpha^i \bar{x}_r(k_0) + \bar{y}_{r,(i)}(k_0)) + 2\bar{\xi}(k_d),$$

where $\bar{y}_{r,(i)}(k_0) \triangleq \max_{x_p \in \mathscr{E}^{x_p}} |y_{r,(i)}(k_0)|$ *and* $\Delta u \triangleq u' - u$ *is the difference between delayed and actual input, then the attack will be detected at the time instant* k_d.

5 Conclusion and Future Work

In this chapter, we introduced a basic physical watermarking scheme where a random noise is injected into the system to excite the system and check whether the system responds to the watermark signal in accordance to the dynamical model of the system. The optimal watermark is derived via solving an optimization problem which aims to achieve the optimal trade-off between control performance and detection performance. Then three interesting extensions about the watermark design were presented in detail.

For future works, it is worth noticing how to apply the watermark scheme to more complicated systems. Designing more efficient algorithms regarding the watermarking signal against more intelligent attackers is also interesting. Also, it is of great interest to test the proposed algorithms in CPS to verify their performance in a real scenario.

References

[1] Slay J, Miller M (2007) Lessons learned from the maroochy water breach. In: International Conference on Critical Infrastructure Protection, Springer, pp 73–82

[2] Karnouskos S (2011) Stuxnet worm impact on industrial cyber-physical system security. In: IECON 2011-37th Annual Conference of the IEEE Industrial Electronics Society, IEEE, pp 4490–4494

[3] Whitehead DE, Owens K, Gammel D, Smith J (2017) Ukraine cyber-induced power outage: Analysis and practical mitigation strategies. In: 2017 70th Annual Conference for Protective Relay Engineers (CPRE), IEEE, pp 1–8

[4] Wikipedia contributors (2019) 2019 venezuelan blackouts — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/w/index.php?title=2019_Venezuelan_blackouts&oldid=908146648`, [Online; accessed 21-August-2019]

[5] Mo Y, Weerakkody S, Sinopoli B (2015) Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor

outputs. IEEE Control Systems Magazine 35(1):93–109
[6] Wikipedia contributors (2019) Digital watermarking — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Digital_watermarking&oldid=910119309, [Online; accessed 22-August-2019]
[7] Mo Y, Sinopoli B (2009) Secure control against replay attacks. In: 2009 47th annual Allerton conference on communication, control, and computing (Allerton), IEEE, pp 911–918
[8] Mo Y, Chabukswar R, Sinopoli B (2013) Detecting integrity attacks on scada systems. IEEE Transactions on Control Systems Technology 22(4):1396–1407
[9] Chabukswar R, Mo Y, Sinopoli B (2011) Detecting integrity attacks on scada systems. IFAC Proceedings Volumes 44(1):11,239–11,244
[10] Khazraei A, Kebriaei H, Salmasi FR (2017) A new watermarking approach for replay attack detection in lqg systems. In: 2017 IEEE 56th Annual Conference on Decision and Control (CDC), IEEE, pp 5143–5148
[11] Khazraei A, Kebriaei H, Salmasi FR (2017) Replay attack detection in a multi agent system using stability analysis and loss effective watermarking. In: 2017 American Control Conference (ACC), IEEE, pp 4778–4783
[12] Weerakkody S, Ozel O, Sinopoli B (2017) A bernoulli-gaussian physical watermark for detecting integrity attacks in control systems. In: 2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, pp 966–973
[13] Satchidanandan B, Kumar PR (2016) Dynamic watermarking: Active defense of networked cyber–physical systems. Proceedings of the IEEE 105(2):219–240
[14] Satchidanandan B, Kumar P (2019) On the design of security-guaranteeing dynamic watermarks. IEEE Control Systems Letters 4(2):307–312
[15] Liu H, Yan J, Mo Y, Johansson KH (2018) An on-line design of physical watermarks. In: 2018 IEEE Conference on Decision and Control (CDC), IEEE, pp 440–445
[16] Rubio-Hernán J, De Cicco L, Garcia-Alfaro J (2016) Revisiting a watermark-based detection scheme to handle cyber-physical attacks. In: 2016 11th International Conference on Availability, Reliability and Security (ARES), IEEE, pp 21–28
[17] Rubio-Hernan J, De Cicco L, Garcia-Alfaro J (2017) On the use of watermark-based schemes to detect cyber-physical attacks. EURASIP Journal on Information Security 2017(1):8
[18] Rubio-Hernan J, De Cicco L, Garcia-Alfaro J (2016) Event-triggered watermarking control to handle cyber-physical integrity attacks. In: Nordic Conference on Secure IT Systems, Springer, pp 3–19
[19] Fang C, Qi Y, Cheng P, Zheng WX (2017) Cost-effective watermark based detector for replay attacks on cyber-physical systems. In: Control Conference (ASCC), 2017 11th Asian, IEEE, pp 940–945
[20] Ferrari RM, Teixeira AM (2017) Detection and isolation of replay attacks through sensor watermarking. IFAC-PapersOnLine 50(1):7363–7368

[21] Ferrari RM, Teixeira AM (2017) Detection and isolation of routing attacks through sensor watermarking. In: 2017 American Control Conference (ACC), IEEE, pp 5436–5442
[22] Teixeira AM, Ferrari RM (2018) Detection of sensor data injection attacks with multiplicative watermarking. In: 2018 European Control Conference (ECC), IEEE, pp 338–343
[23] Mehra RK, Peschon J (1971) An innovations approach to fault detection and diagnosis in dynamic systems. Automatica 7(5):637–640
[24] Teixeira A, Shames I, Sandberg H, Johansson KH (2012) Revealing stealthy attacks in control systems. In: 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), IEEE, pp 1806–1813
[25] Hoehn A, Zhang P (2016) Detection of covert attacks and zero dynamics attacks in cyber-physical systems. In: 2016 American Control Conference (ACC), IEEE, pp 302–307
[26] Mo Y, Sinopoli B (2010) False data injection attacks in control systems. In: Preprints of the 1st workshop on Secure Control Systems, pp 1–6
[27] Liu Y, Ning P, Reiter MK (2011) False data injection attacks against state estimation in electric power grids. ACM Transactions on Information and System Security (TISSEC) 14(1):13
[28] Scharf LL (1991) Statistical signal processing, vol 98. Addison-Wesley Reading, MA
[29] Chonavel T (2002) Statistical signal processing: modelling and estimation. Springer Science & Business Media
[30] Delsarte P, Genin Y, Kamp Y (1978) Orthogonal polynomial matrices on the unit circle. IEEE Transactions on Circuits and Systems 25(3):149–160
[31] Ferrari RM, Parisini T, Polycarpou MM (2008) A robust fault detection and isolation scheme for a class of uncertain input-output discrete-time nonlinear systems. In: 2008 American Control Conference, IEEE, pp 2804–2809