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

Switching Watermarking-based Detection Scheme Against Replay Attacks

Lijing Zhai[1], Kyriakos G. Vamvoudakis[1], Jérôme Hugues[2]

2021 60th IEEE Conference on Decision and Control (CDC) | 978-1-6654-3659-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CDC45484.2021.9683652

Abstract— **In this paper, a switching watermarking-based detection scheme is proposed to detect replay attacks while also limiting the adversary's knowledge about additive watermarking signals to ensure unpredictability. The unpredictability is introduced by appropriate switching through a trade-off between detection performance and information entropy, which makes it challenging for adversaries to estimate the additive watermarking signals. In addition, we compare the detection performance of Neyman-Pearson detector and χ^2 detector. Simulation results show the efficacy of the proposed approach.**

Index Terms— **CPS, watermarking signals, replay attacks, detector, switching.**

I. Introduction

Cyber-physical systems (CPS) are composed of physical devices with computational and communication components. The integration and interaction between cyber and physical components add automation capabilities that improve physical systems and processes in various critical domains. However, CPS complexity and heterogeneity lead to their vulnerability to be exploited by malicious adversaries [1]. A tremendous amount of research have been conducted to design resilient CPS, i.e., increasing their abilities of detecting and mitigating malicious attacks [2].

In this work, we focus on replay attacks, which record sensory output signals for a period of time and then replay the recorded values at a later time. This makes it difficult for the system operators to distinguish between nominal output signals and replayed output signals. Research efforts have been made to utilize physical watermarking to defend against replay attacks. The key idea of watermarking-based detection scheme is to superimpose optimal control input signals with watermarking signals that serve as an authentication for the detection of malicious attacks. The additive watermarking signals into the system deteriorate the control performance, and thus an optimization problem of maximizing detection performance with a constraint of certain maximal control performance loss is formulated to generate optimal watermarking signals [3], [4].

Based on the basic physical watermarking signals, several extensions of the watermarking generation schemes are also proposed, such as: dynamic watermarking scheme [5], [6], data-based watermarking scheme [7], [8], and multiplicative watermarking scheme [9], [10]. The work of [11] developed a dynamic watermarking approach to detect malicious sensor attacks for general LTI systems with partial state observations, and proposed an internal model principle-based approach to handle persistent disturbances. The work of [12] incorporated Bernoulli packet drops at the control input in the design of the watermarking signals and developed a joint Bernoulli-Gaussian watermarking scheme with a correlation detector to detect integrity attacks. The authors in [13] developed a periodic watermarking strategy for detecting discontinuous replay attacks to reduce the control cost.

In terms of the detector schemes, various statistical tests are investigated, such as χ^2 detection scheme using innovation data [3], [14], Neyman-Pearson detection scheme using observation data [7], cumulative sum detection scheme using both innovation signals and watermarking signals [15]. The authors of [16]–[18] categorized adversary models as cyber adversaries (i.e., attacks are able to eavesdrop sensory output from the system without acquiring system dynamics and to inject arbitrary signals at the sensors to conduct malicious actions), non-parametric cyber-physical adversaries (i.e., attacks are able to eavesdrop input and output information from the system to estimate system dynamics and to inject malicious signals) and parametric cyber-physical adversaries (i.e., attacks are able to use the estimated parameters of the system from input and output data to mislead the controller).

The work of [19] discussed two methods of active detection in CPS namely, physical watermarking and moving target defense (MTD). Different from physical watermarking scheme, MTD defense scheme introduces time-varying and unpredictable properties to keep adversaries unaware of the full system model. Motivated by this concept, in this work we incorporate the idea of MTD into the physical watermarking detection scheme. Instead of switching system dynamics, we consider switching watermarking generation. In particular, we extend our previous work [8] to develop a switching watermarking-based detection scheme that incorporates unpredictability into the watermarking generation to build a robust watermarking-detection framework.

Contributions: Compared with the deterministic multiple-watermarking detection scheme in the literature, the scheme developed in this work adds unpredictability in terms of switching, which makes it difficult for intelligent adversaries to estimate additive watermarking signals. In addition, we compare the detection performance between

[1] L. Zhai and K. G. Vamvoudakis are with the Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA e-mail: (lzhai3@gatech.edu, kyriakos@gatech.edu).

[2] J. Hugues is with the Carnegie Mellon University/Software Engineering Institute, Pittsburgh, PA 15213, USA e-mail: jhugues@andrew.cmu.edu.

This work was supported in part by the Department of Energy under grant No. DE-EE0008453, by ONR Minerva under grant No. N00014 − 18 − 1 − 2160, by NSF under grant Nos. CAREER CPS-1851588 and S&AS 1849198, and by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. DM21-0275.

Neyman-Pearson detector and χ^2 detector against attacks that are able to estimate system dynamics from input and output data and to inject malicious input into the system during the attack period.

Structure: Section II formulates the problem and defines attack strategy. The designs of Neyman-Pearson detector, optimal watermarking signals and a switching watermarking-based detection scheme are shown in Section III. Section IV presents simulation results and finally Section V concludes and talks about future work.

II. Problem Formulation

A. System Model

Consider the following LTI discrete-time system $\forall k \in \mathbb{Z}$,

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

where $x_k \in \mathbb{R}^n$ is the state, $u_k \in \mathbb{R}^l$ is the control input, $y_k \in \mathbb{R}^m$ is the output, and $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times l}$, and $C \in \mathbb{R}^{m\times n}$ are the state matrix, control input matrix, and output matrix, respectively. In this work, the process noise $w_k \in \mathbb{R}^n$ is assumed to be a zero-mean Gaussian noise with covariance $\Sigma_w > 0$ denoted as $w_k \sim \mathcal{N}(0, \Sigma_w)$, and the measurement noise $v_k \in \mathbb{R}^m$ is assumed to follow a zero-mean Gaussian distribution with covariance $\Sigma_v > 0$ denoted as $v_k \sim \mathcal{N}(0, \Sigma_v)$. In addition, the initial condition x_0 is assumed to follow a zero-mean Gaussian distribution with covariance Σ_0 denoted as $x_0 \sim \mathcal{N}(0, \Sigma_0)$. The signals w_k and v_k are uncorrelated. The initial condition x_0 is independent of process and measurement noises.

Assumption 1. Since CPS operate for an extended period of time, we assume they are operating at a steady state. □

Assumption 2. Assume that the pairs (A, B) and (A, C) are controllable and observable respectively. □

Define the cost functional as $J = \mathbf{E}[\sum_{i=0}^{\infty}\gamma^i(y_i^{\mathrm{T}}Qy_i + u_i^{\mathrm{T}}Ru_i)]$, where $Q \geq 0 \in \mathbb{R}^{m\times m}$, $R > 0 \in \mathbb{R}^{l\times l}$, $0 < \gamma < 1$ is a discount factor, and $\mathbf{E}$ denotes expectation operator. Given (1)-(2), we aim to find a controller that minimizes the cost functional with the following optimal value $\forall x_k$,

$$V^\star(x_k) = \min_{u_k} \mathbf{E}[\sum_{i=k}^{\infty} \gamma^{i-k}(y_i^{\mathrm{T}}Qy_i + u_i^{\mathrm{T}}Ru_i)]. \tag{3}$$

Since information about full states is not always accessible for feedback, we shall design an observer $\forall k \in \mathbb{Z}$ as,

$$\hat{x}_{k+1} = A\hat{x}_k + Bu_k + Ke_k, \tag{4}$$
$$y_k = C\hat{x}_k + e_k, \tag{5}$$

where $\hat{x}_k \in \mathbb{R}^n$ denotes the estimated state and K stands for the steady-state Kalman filter gain while the term $e_k \in \mathbb{R}^m$ is known as the innovation with covariance $E_k = \mathrm{cov}(e_k) = \mathbf{E}(e_k e_k^{\mathrm{T}})$, where $\mathbf{E}(\cdot)$ denotes expectation operator. The inovation E_k is calculated as $E_k = CSC^{\mathrm{T}} + \Sigma_v$. The steady-state Kalman gain is computed as $K = ASC^{\mathrm{T}}(CSC^{\mathrm{T}} + \Sigma_v)^{-1}$, where S stands for the covariance of the estimation error $\varepsilon_k := x_k - \hat{x}_k$. The covariance of the estimation error $\mathrm{cov}(\varepsilon_k) = \mathbf{E}(\varepsilon_k\varepsilon_k^{\mathrm{T}})$ is calculated by solving the algebraic Ricatti equation (ARE) of $S = ASA^{\mathrm{T}} + \Sigma_w - ASC^{\mathrm{T}}\left(CSC^{\mathrm{T}} + \Sigma_v\right)^{-1}CSA^{\mathrm{T}}$. We can obtain the covariance of the estimated state $X_k = \mathrm{cov}(\hat{x}_k) = \mathbf{E}(\hat{x}_k\hat{x}_k^{\mathrm{T}})$ by solving the following ARE $\forall k \in \mathbb{Z}$,

$$X_k = AX_kA^{\mathrm{T}} + ASC^{\mathrm{T}}\left(CSC^{\mathrm{T}} + \Sigma_v\right)^{-1}CSA^{\mathrm{T}}. \tag{6}$$

The optimal control law $\forall k \in \mathbb{Z}$ is,

$$u_k^\star = L\hat{x}_k, \ \forall \hat{x}_k, \tag{7}$$

with the feedback gain $L = -(B^{\mathrm{T}}PB + R/\gamma)^{-1}B^{\mathrm{T}}PA$, where P is found by solving the ARE of $P = \gamma A^{\mathrm{T}}PA + C^{\mathrm{T}}QC - \gamma A^{\mathrm{T}}PB(B^{\mathrm{T}}PB + R/\gamma)^{-1}B^{\mathrm{T}}PA$.

B. Replay Attack Strategy

Definition 1. Replay attacks considered in this work are defined as follows.

- Step 1. Adversaries record a sequence of sensory measurements from k_1 to k_1+T, where $T \in \mathbb{Z}^+$ is chosen by adversaries to be large enough to guarantee the sequence can be replayed for an extended period of time.
- Step 2. Adversaries replace the current sensory measurements y_k with the recorded ones, i.e., $y_k' = y_{k-\Delta k}$, from k_2 to $k_2 + T$, where y_k' denotes replayed signals and $\Delta k := k_2 - k_1$. □

III. Watermarking-based Detection Scheme

In order to actively defend against replay attacks, physical watermarking signals ϕ_k are injected into the control input to serve as an authentication. Thus, the overall control input u_k of the system is the summation of the optimal control input and watermarking signals given by $u_k = u_k^\star + \phi_k$. It is assumed that watermarking signals ϕ_k follow a zero-mean Gaussian distribution with covariance $U > 0$, i.e., $\phi_k \sim \mathcal{N}(0, U)$. By adding watermarking signals to the system, the distributions of measurement output without replay attacks and under replay attacks will be different, and thus a detector can be designed due to such a statistical difference.

A. Neyman-Pearson Detector

Next, we shall characterize the distributions of the output data, with and without considering replay attacks. Given the system dynamics (1) and (2), it can be shown that the output signals without replay attacks $\forall k \in \mathbb{Z}$ are,

$$\begin{aligned} y_k &= \sum_{t=0}^{k-1} CA^tB\phi_{k-1-t} + \sum_{t=0}^{k-1} CA^tBu_{k-1-t}^\star \\ &\quad + \sum_{t=0}^{k-1} CA^t w_{k-1-t} + v_k + CA^k x_0 \\ &= \underbrace{\sum_{t=0}^{k-1} CA^tB\phi_{k-1-t}}_{\text{first term}} + \underbrace{\sum_{t=0}^{k-1} CA^tBL\hat{x}_{k-1-t}}_{\text{second term}} \end{aligned}$$

$$+\underbrace{\sum_{t=0}^{k-1} CA^t w_{k-1-t} + v_k + CA^k x_0}_{\text{third term}}.$$

For simplicity, we will split y_k into three terms and denote them respectively $\forall k \in \mathbb{Z}$ as,

$$\Phi_k := \sum_{t=0}^{k} CA^t B\phi_{k-t}, \tag{8}$$

$$\varrho_k := \sum_{t=0}^{k} CA^t BL\hat{x}_{k-t}, \tag{9}$$

$$\theta_k := \sum_{t=0}^{k} CA^t w_{k-t} + v_{k+1} + CA^{k+1}x_0. \tag{10}$$

Thus, the measurement output $\forall k \in \mathbb{Z}$ is characterized by,

$$y_k = \Phi_{k-1} + \varrho_{k-1} + \theta_{k-1}. \tag{11}$$

From (8), we can see that Φ_{k-1} is a zero-mean Gaussian distribution with covariance given by,

$$\Upsilon := \sum_{\tau=0}^{\infty} (CA^\tau B)U(CA^\tau B)^{\mathrm{T}}. \tag{12}$$

Similarly, it can be observed from (9) that ϱ_{k-1} is a zero-mean Gaussian distribution with covariance given by,

$$\Gamma := \sum_{\tau=0}^{\infty} [CA^\tau BL]X_k[CA^\tau BL]^{\mathrm{T}}. \tag{13}$$

By observing (10) we can see that θ_{k-1} is a zero-mean Gaussian distribution with covariance given by $\Theta := C\Sigma C^{\mathrm{T}} + \Sigma_v$, where Σ stands for the covariance of state x_k for the dynamical system without control input, i.e., $x_{k+1} = Ax_k + w_k,\ y_k = Cx_k + v_k$, and Σ satisfies $\Sigma = A\Sigma A^{\mathrm{T}} + \Sigma_w$.

In the case that the system is under replay attacks, the replayed output is described $\forall k \in \mathbb{Z}$ as,

$$y_k' = y_{k-\Delta k} = \Phi_{k-1-\Delta k} + \varrho_{k-1-\Delta k} + \theta_{k-1-\Delta k}. \tag{14}$$

From (11) we can conclude that in the nominal (attack-free) case, given collected data $\phi_0, \phi_1, \ldots, \phi_{k-1}$, and $u_0, u_1, \ldots, u_{k-1}$, the measurement output y_k converges to a Gaussian distribution with mean $\Phi_{k-1} + \varrho_{k-1}$ and covariance Θ, denoted as $y_k \sim \mathcal{N}_0(\Phi_{k-1} + \varrho_{k-1}, \Theta)$. In the case that the system is under replay attacks, the parameter Δk is unknown to the system operator even though it is deterministic. Therefore, from (14) we can conclude that the output y_k' converges to a zero-mean Gaussian distribution with covariance $\Upsilon + \Gamma + \Theta$, denoted as $y_k' \sim \mathcal{N}_1(0, \Upsilon + \Gamma + \Theta)$.

Since y_k and y_k' follow two different Gaussian distributions, we can design a detector to distinguish this statistical difference. We will now consider the following binary hypothesis testing with the null hypothesis given by $\mathcal{H}_0$ and the alternative hypothesis given by $\mathcal{H}_1$.

- $\mathcal{H}_0$: The measurement output y_k follows a Gaussian distribution $\mathcal{N}_0(\Phi_{k-1} + \varrho_{k-1}, \Theta)$.
- $\mathcal{H}_1$: The measurement output y_k' follows a Gaussian distribution $\mathcal{N}_1(0, \Upsilon + \Gamma + \Theta)$.

Based on Neyman-Pearson lemma [20], the alarm signal of Neyman-Pearson detector for the hypothesis $\mathcal{H}_0$ versus the hypothesis $\mathcal{H}_1$ is calculated $\forall k \in \mathbb{Z}$ as,

$$\begin{aligned} g_k =& (y_k - \Phi_{k-1} - \varrho_{k-1})^{\mathrm{T}}\Theta^{-1}(y_k - \Phi_{k-1} - \varrho_{k-1}) \\ &- y_k^{\mathrm{T}}(\Upsilon + \Gamma + \Theta)^{-1}y_k. \end{aligned} \tag{15}$$

Then the alarm signal g_k is compared to a predetermined threshold ξ which is tuned based on the detection performance, such as detection and false alarm rate. Note that, $g_k \geqslant \xi$ implies that the hypothesis $\mathcal{H}_1$ is valid and thus the system is under replay attacks. Otherwise, the null hypothesis $\mathcal{H}_0$ holds with $g_k < \xi$ and the system operates normally.

Definition 2. We define intelligent attacks as follows.

- Adversaries record a sequence of sensory measurements from k_1 to $k_1 + T$ with $T \in \mathbb{Z}^+$.
- Adversaries replace current sensory measurements y_k with recorded ones, i.e., $y_k' = y_{k-\Delta k}$, from k_2 to $k_2 + T$ with $T \in \mathbb{Z}^+$.
- During the interval $[k_2, k_2 + T]$, $T \in \mathbb{Z}^+$, adversaries can inject malicious input Bu_k^a into the system. □

Remark 1. The attacks defined in Definition 2 are able to use estimated parameters of the system based on input and output data to estimate additive watermarking signals.

Theorem 1. Suppose the system (1)-(2) is compromised by intelligent attacks of Definition 2, then Neyman-Pearson detector with alarm signals (15) can detect these attacks.

Proof. Suppose attacks start at k. The measurement output received by the controller is y_k' given by (14). At the same time, the attacks inject malicious input Bu_k^a into the system. As a result, the system under attacks is described $\forall k \in \mathbb{Z}$ by $x_{k+1} = Ax_k + Bu_k + Bu_k^a + w_k$, which shows that the injected malicious input Bu_k^a influences the state trajectory from time $k+1$ thereafter. Note that Neyman-Pearson detector utilizes measurement output data to calculate alarm signals instead of innovation signals (i.e., the difference between measured and estimated output) used by χ^2 detector, and that replay attackers replay the previous output data $y_{k-\Delta k}$ which are not influenced by attackers' ability to inject malicious input during the active attack period defined by Step 2 in Definition 1. Thus, the difference of distributions between case of no attacks $\mathcal{N}_0$ and case with attacks $\mathcal{N}_1$ still exists and the alarm signal given by (15) is effective. ■

Next, Kullback-Liebler (KL) divergence is utilized to characterize the Neyman-Pearson detector performance [3].

Lemma 1. The expectation of KL-divergence for distributions $\mathcal{N}_0$ and $\mathcal{N}_1$ is given by,

$$\begin{aligned} &\mathbf{E}[D_{\mathrm{KL}}(\mathcal{N}_0, \mathcal{N}_1)] \\ =&\mathrm{tr}[(\Upsilon + \Gamma)\Theta^{-1}] - \frac{1}{2}\log\{\det[I_m + (\Upsilon + \Gamma)\Theta^{-1}]\}, \end{aligned} \tag{16}$$

and its lower bound and upper bound are given by,

$$\frac{1}{2}\mathrm{tr}[(\Upsilon + \Gamma)\Theta^{-1}] \leqslant \mathbf{E}[D_{\mathrm{KL}}(\mathcal{N}_0, \mathcal{N}_1)] \leqslant \mathrm{tr}[(\Upsilon + \Gamma)\Theta^{-1}]$$

$$-\frac{1}{2}\log\{1+\mathrm{tr}[(\Upsilon+\Gamma)\Theta^{-1}]\},$$

where tr$(\cdot)$ denotes the trace operator [8].

B. Optimal Watermarking Signals

Although replay attacks can be detected with the help of additive watermarking signals, the system performance is deteriorated by deviating from optimal control law. Next we quantify the performance loss.

Lemma 2. Given the system (1)-(2) with the optimal control law (7), the performance loss due to additive watermarking signals with zero mean and covariance U is given by $\Delta J = \mathrm{tr}(UR)/(1-\gamma)$.

Proof. For the measurement signals defined by (11), the performance cost without considering additive watermarking signals is given by $J^\star = \mathbf{E}[\sum_{i=0}^{\infty}\gamma^i(y_i^{\mathrm{T}}Qy_i + (u_i^\star)^{\mathrm{T}}Ru_i^\star)]$. With the consideration of additive watermarking signals, the performance changes to,

$$\begin{aligned}J &= \mathbf{E}\{\sum_{i=0}^{\infty}\gamma^i[y_i^{\mathrm{T}}Qy_i + (u_i^\star+\phi_i)^{\mathrm{T}}R(u_i^\star+\phi_i)]\}\\ &= J^\star + \mathbf{E}[\sum_{i=0}^{\infty}\gamma^i((u_i^\star)^{\mathrm{T}}R\phi_i + \phi_i^{\mathrm{T}}Ru_i^\star + \phi_i^{\mathrm{T}}R\phi_i)]\\ &= J^\star + \sum_{i=0}^{\infty}\gamma^i\mathrm{tr}(UR) = J^\star + \frac{\mathrm{tr}(UR)}{1-\gamma},\end{aligned}$$

which completes the proof. ■

Remark 2. The control performance loss ΔJ due to additive watermarking signals is linearly dependent on covariance U. But as shown in (16) the detection performance characterized by the KL-divergence is not linearly dependent on U. □

Next, we aim to find the trade-off between the detection and control performances. Note that both the upper and lower bounds of the KL-divergence in (16) contain tr$[(\Upsilon+\Gamma)\Theta^{-1}]$, thus we formulate the following optimization problem,

$$\begin{aligned}U = \arg\max_{U} \quad & \mathrm{tr}[(\Upsilon+\Gamma)\Theta^{-1}]\\ \text{subject to} \quad & \mathrm{tr}(UR)\leqslant\zeta,\end{aligned}$$

where $\zeta\in\mathbb{R}^+$ is a threshold determining how much control performance loss can be tolerated. Substituting covariance expressions (12) and (13) into tr$[(\Upsilon+\Gamma)\Theta^{-1}]$ yields,

$$\begin{aligned}\mathrm{tr}[(\Upsilon+\Gamma)\Theta^{-1}] =& \mathrm{tr}[\sum_{\tau=0}^{\infty}(H_\tau UH_\tau^{\mathrm{T}} + \Omega_\tau X_k\Omega_\tau^{\mathrm{T}})\Theta^{-1}]\\ =& \mathrm{tr}(U\mathcal{P}_1) + \mathrm{tr}(X_k\mathcal{P}_2),\end{aligned}$$

with $H_\tau \triangleq CA^\tau B, \Omega_\tau \triangleq CA^\tau BL, \mathcal{P}_1 \triangleq \sum_{\tau=0}^{\infty}H_\tau^{\mathrm{T}}\Theta^{-1}H_\tau$, and $\mathcal{P}_2 \triangleq \sum_{\tau=0}^{\infty}\Omega_\tau^{\mathrm{T}}\Theta^{-1}\Omega_\tau$.

Note that $\mathcal{P}_1$ and $\mathcal{P}_2$ are related to the system dynamics (4)-(5), and that X_k can be found by solving the Lyapunov equation (6). These three parameters are not related to the covariance of watermarking signals U. It is also true that tr$(X_k\mathcal{P}_2)\geqslant 0$ always holds since X_k is positive semi-definite. Therefore, we can optimize the KL-divergence by maximizing tr$(U\mathcal{P}_1)$. As a result, we can solve instead the following optimization problem to get the covariance of optimal watermarking signals,

$$\begin{aligned}U = \arg\max_{U} \quad & \mathrm{tr}(U\mathcal{P}_1)\\ \text{subject to} \quad & \mathrm{tr}(UR)\leqslant\zeta,\end{aligned} \tag{17}$$

with $\mathcal{P}_1 = \sum_{\tau=0}^{\infty}(CA^\tau B)^{\mathrm{T}}\Theta^{-1}(CA^\tau B)$.

Theorem 2. The solution to the optimization problem (17) is given by $U = zz^{\mathrm{T}}$, with z as the eigenvector corresponding to the largest eigenvalue of the matrix $R^{-1}\mathcal{P}_1$ and satisfying $z^{\mathrm{T}}Rz = \zeta$ [8].

C. Switching Watermarking Detection Scheme

As stated in [16]–[18], resourceful and powerful attackers can employ an adaptive least mean square filter to identify system model with input and output data and thus estimate the watermarking signals added into the system. By switching the distributions when generating watermarking signals, defenders can introduce time-varying and unpredictable properties as a moving target to increase attackers' difficulty in accessing to the additive watermarking signals, and thus improve the probability of detecting replay attacks.

Generally speaking, the idea of the switching watermarking-based detection scheme is to generate additive watermarking signals by periodically switching among N different Gaussian distribution modes. Specifically, $\phi_k^{(i)}$ stands for additive watermarking signals at the time instant k generated by solving the optimization problem (17) with control performance loss threshold ζ_i for the i-th Gaussian distribution mode, where $i\in\mathcal{I}=\{1,2,\ldots,N\}$. From Theorem 2, the covariance of the optimal watermarking signals by solving the optimization problem (17) only depends on the control performance loss threshold ζ and thus various distributions can be generated with different thresholds. At each switching moment, a distribution mode i is selected randomly with a switching period T.

This watermarking generation scheme adds unpredictability by switching but the detection performance is degraded since the distribution with the maximum KL-divergence standing for the best detection performance is not always utilized. Thus, a switching law can be developed by finding the trade-off between detection performance characterized by KL-divergence (16) and unpredictability characterized by information entropy, i.e., $\mathcal{H}(\mathbf{p}) = -\mathbf{p}^{\mathrm{T}}\log(\mathbf{p})$ [21], where $\mathbf{p}=\{p_1,p_2,\ldots,p_N\}$ is the probability simplex, satisfying $\|\mathbf{p}\|_1 = \sum_{i=1}^{N}|p_i| = 1$ and standing for the selected probability of each distribution mode at the switching moment.

Lemma 3. Suppose additive watermarking signals are generated by periodically switching among N different Gaussian distribution modes and the covariance of the optimal watermarking signals is calculated by solving the optimization problem (17) with control performance loss threshold ζ_i for each mode, where $i\in\mathcal{I}=\{1,2,\ldots,N\}$. Then the probability simplex is solved by,

$$\min_{\mathbf{p}}(-\mathbf{D_{KL}}^{\mathrm{T}}\mathbf{p} - \epsilon\mathcal{H}(\mathbf{p}))$$

$$\text{such that } \|\mathbf{p}\|_1 = 1, p_i \geqslant 0, i \in \{1, 2, \dots, N\}, \tag{18}$$

and the solution is given by,

$$p_i = e^{[\frac{D_{\mathrm{KL}}^{(i)}}{\epsilon} - 1 - \log(e^{-1} \sum_{i=1}^{N} e^{\frac{D_{\mathrm{KL}}^{(i)}}{\epsilon}})]}, \tag{19}$$

where ϵ denotes the weight on unpredictability, $\mathbf{D_{KL}} = [D_{\mathrm{KL}}^{(1)};\ D_{\mathrm{KL}}^{(1)};\ \dots;\ D_{\mathrm{KL}}^{(N)}]$ denotes the KL-divergence of each distribution calculated by equation (16).

Proof. Define the Lagrangian for (18) as $L = -\mathbf{D_{KL}}^{\mathrm{T}}\mathbf{p} - \epsilon\mathcal{H}(\mathbf{p}) + \lambda(\mathbf{1}^{\mathrm{T}}\mathbf{p} - 1) + \beta^{\mathrm{T}}\mathbf{p} = -\mathbf{D_{KL}}^{\mathrm{T}}\mathbf{p} + \epsilon\mathbf{p}^{\mathrm{T}}\log(\mathbf{p}) + \lambda(\mathbf{1}^{\mathrm{T}}\mathbf{p} - 1) + \beta^{\mathrm{T}}\mathbf{p}$. Apply KKT conditions then we get $\Delta_{\mathbf{p}}L = -\mathbf{D_{KL}} + \epsilon\log(\mathbf{p}) + \epsilon\mathbf{1} + \lambda\mathbf{1} + \beta = 0$ and $\beta^{\mathrm{T}}\mathbf{p}^{\star} = 0$. Considering the feasibility of log operation and the nontrivial situation ($\beta = \mathbf{0}$), it follows that $\Delta_{\mathbf{p}}L = -\mathbf{D_{KL}} + \epsilon\log(\mathbf{p}) + \epsilon\mathbf{1} + \lambda\mathbf{1} = 0$. Then for each $i \in \{1, 2, \dots, N\}$, we get,

$$p_i = e^{\frac{D_{\mathrm{KL}}^{(i)}}{\epsilon} - \frac{\lambda}{\epsilon} - 1}. \tag{20}$$

Consider constraint $\|\mathbf{p}\|_1 = 1$. Solve for λ from (20),

$$\lambda = \epsilon\log(e^{-1}\Sigma_{i=1}^{N} e^{\frac{D_{\mathrm{KL}}^{(i)}}{\epsilon}}). \tag{21}$$

Substitute (21) in (20) to get the required result. ■

The pseudocode that describes the proposed switching watermarking-based detection scheme against replay attacks is summarized as Algorithm 1.

Algorithm 1 Switching Watermarking-based Detection Scheme Against Replay Attacks

```
Procedure
    Given $N$ different control loss thresholds $\zeta_i$, where $i \in \{1, 2, \dots, N\}$.
    for $i = \{1, 2, \dots, N\}$
        Solve the optimization problem (17) to get the covariance of optimal watermarking signals $U^i$ for $\zeta_i$.
        Compute the KL-divergence $D_{\mathrm{KL}}^{(i)}$ using (16).
        Generate watermarking signals following the Gaussian distribution $\phi_k^i \sim \mathcal{N}(0, U^i)$.
    end for
    Solve for the probability simplex $\mathbf{p}$ using (19).
    At $k = 0$, choose the best distribution mode, i.e., $\sigma(0) = \arg\max_{i \in \mathcal{I}}(p_i)$.
    Propagate the system according to (1) and (2).
    Compute the optimal control input $u_k^{\star}$ using (7).
    Switch to another distribution based on the probability simplex $\mathbf{p}$ at the switching moment with switching period $T$.
    Compute alarm signals $g_k$ using (15).
    Raise alarms when $g_k \geqslant \xi$, where $\xi$ is the alarm threshold.
End procedure
```

IV. Simulation

Consider the following system $\forall k \in \mathbb{Z}$,

$$\begin{aligned} x_{k+1} &= \begin{bmatrix} 1.1 & -0.3 \\ 1 & 0 \end{bmatrix} x_k + \begin{bmatrix} 1 \\ 0 \end{bmatrix} (u_k^{\star} + \phi_k) + w_k, \\ y_k &= \begin{bmatrix} 1 & -0.8 \end{bmatrix} x_k + v_k, \end{aligned}$$

where $u_k^{\star}$ is the optimal control given by (7). We choose Q and R in (3) as identity matrices of appropriate dimensions, and $\gamma = 0.8$. Assume that the system is subjected to a zero-mean white noise with covariances $\Sigma_w = 0.06$ and $\Sigma_v = 0.06$. The probability simplex solved by (19) is $\mathbf{p} = [0.011;\ 0.647;\ 0.231;\ 0.029;\ 0.082]$ with $\zeta = \{1.2,\ 1.6,\ 1.4,\ 1.3,\ 1.5\}$. This shows the second and third Gaussian distribution modes have better detection performance than other modes. With the running time 400 seconds and switching period $T = 8$ seconds, a replay attacker records sensory output during 101-200 seconds and then replays the recorded data during 201-300 seconds.

Figure 1 shows the evolution of system states, measurement output, optimal control input $u_k^{\star}$ and optimal watermarking signals ϕ_k. The alarm signals of the Neyman-Pearson detector are shown in Figure 2. It can be seen the alarm signals have a great increase from 201 second to 300 second implying the existence of replay attacks. The evolution of switching signals is shown in Figure 3. The probabilities of the second and third distribution modes are the biggest and thus Figure 3 shows these two modes are selected more frequently than other modes.

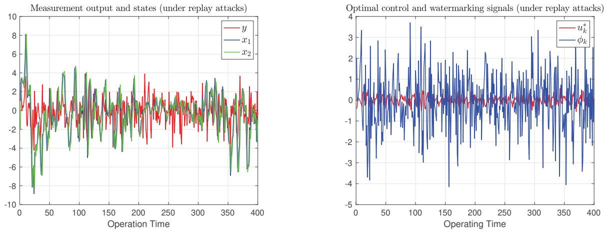

Fig. 1. Evolution of system states, measurement output, optimal control input $u_k^{\star}$ and optimal watermarking signals ϕ_k. Replay attacks record sensory output during 101-200 seconds and then replay such recorded data during 201-300 seconds.

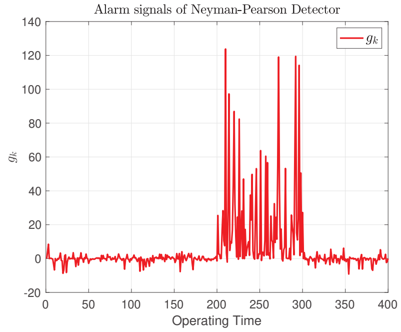

Fig. 2. Evolution of the alarm signals g_k under replay attacks of Definition 1. Replay attacks record sensory output during 101-200 seconds and then replay such recorded data during 201-300 seconds.

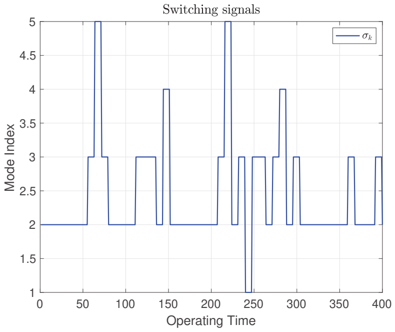

Fig. 3. Evolution of the switching signals. Switching period is $T = 8$ seconds. The second and third distribution modes are selected more frequently than other modes.

Next, we consider systems under attacks of Definition 2. We employ the same simulation setting. Attackers record the measurement output from 51 second to 100 second, then replay these recorded data and inject the signals $-B\phi_k^i$ into the system during the interval of 101-150 seconds. Later the attackers record the measurement output during the interval 201-250 seconds, then replay these recorded data and inject the signals $-B\phi_k^i$ into the system during the interval of 251-300 seconds. This attack setting is the worst-case scenario in practice since attackers can estimate the additive watermarking signals perfectly and add negative watermarking signals to cancel out the addtive watermarking signals. The alarm signals of Neyman-Pearson detector and χ^2 detector are shown in Figure 4, which implies that the Neyman-Pearson detector can detect intelligent attacks of Definition 2 while the χ^2 detector fails.

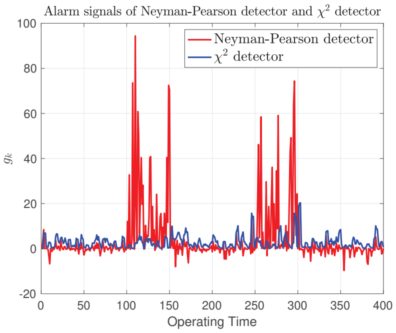

Fig. 4. Evolution of the alarm signals g_k under intelligent attacks of Definition 2. The attackers record the measurement output during 51-100 (resp.201-250) seconds, then replay the recorded data and inject the signals $-B\phi_k$ into the system during 101-150 (resp.251-300) seconds.

V. Conclusion and Future Work

In this work, we develop a switching watermarking-based detection scheme, which adds unpredictability in terms of switching by solving a probability simplex through a trade-off between detection performance and information entropy. Additionally, the comparison of the detection performance between Neyman-Pearson detector and χ^2 detector against intelligent attacks that are able to estimate system dynamics and to inject malicious input signals shows the advantage of Neyman-Pearson detector over χ^2 detector.

References

[1] D. I. Urbina, J. A. Giraldo, A. A. Cardenas, N. O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, and H. Sandberg, "Limiting the impact of stealthy attacks on industrial control systems," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 1092–1105.

[2] S. M. Dibaji, M. Pirani, D. B. Flamholz, A. M. Annaswamy, K. H. Johansson, and A. Chakrabortty, "A systems and control perspective of cps security," *Annual Reviews in Control*, vol. 47, pp. 394–411, 2019.

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

[4] H. Liu, Y. Mo, and K. H. Johansson, "Active detection against replay attack: A survey on watermark design for cyber-physical systems," in *A. M. H. Teixeira and R. M. G. Ferrari, Eds., Safety, Security, and Privacy for Cyber-Physical Systems*. Springer, 2020.

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

[6] S. Weerakkody, O. Ozel, P. Griffioen, and B. Sinopoli, "Active detection for exposing intelligent attacks in control systems," in *2017 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE, 2017, pp. 1306–1312.

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

[8] L. Zhai and K. G. Vamvoudakis, "A data-based private learning framework for enhanced security against replay attacks in cyber-physical systems," *International Journal of Robust and Nonlinear Control*, vol. 31, no. 6, pp. 1817–1833, 2021.

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

[10] ——, "A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks," *IEEE Transactions on Automatic Control*, 2020.

[11] P. Hespanhol, M. Porter, R. Vasudevan, and A. Aswani, "Dynamic watermarking for general lti systems," in *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. IEEE, 2017, pp. 1834–1839.

[12] S. Weerakkody, O. Ozel, and B. Sinopoli, "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, 2017, pp. 966–973.

[13] C. Fang, Y. Qi, P. Cheng, and W. X. Zheng, "Optimal periodic watermarking schedule for replay attack detection in cyber–physical systems," *Automatica*, vol. 112, p. 108698, 2020.

[14] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *2009 47th annual Allerton conference on communication, control, and computing (Allerton)*. IEEE, 2009, pp. 911–918.

[15] A. Naha, A. Teixeira, A. Ahlen, and S. Dey, "Sequential detection of replay attacks," *arXiv preprint arXiv:2012.10748*, 2020.

[16] J. Rubio-Hernán, L. De Cicco, and J. Garcia-Alfaro, "Revisiting a watermark-based detection scheme to handle cyber-physical attacks," in *2016 11th International Conference on Availability, Reliability and Security (ARES)*. IEEE, 2016, pp. 21–28.

[17] J. Rubio-Hernan, L. De Cicco, and J. Garcia-Alfaro, "Event-triggered watermarking control to handle cyber-physical integrity attacks," in *Nordic Conference on Secure IT Systems*. Springer, 2016, pp. 3–19.

[18] ——, "On the use of watermark-based schemes to detect cyber-physical attacks," *EURASIP Journal on Information Security*, vol. 2017, no. 1, pp. 1–25, 2017.

[19] P. Griffioen, S. Weerakkody, B. Sinopoli, O. Ozel, and Y. Mo, "A tutorial on detecting security attacks on cyber-physical systems," in *2019 18th European Control Conference (ECC)*. IEEE, 2019, pp. 979–984.

[20] L. L. Scharf and C. Demeure, *Statistical signal processing: detection, estimation, and time series analysis*. Prentice Hall, 1991.

[21] L. Zhai and K. G. Vamvoudakis, "Data-based and secure switched cyber–physical systems," *Systems & Control Letters*, vol. 148, p. 104826, 2021.