Detection Against Replay Attack: A Feedback Watermark Approach

Xudong Zhao[1(✉)], Le Liu[1], H. R. Karimi[2], and Wei Xing[1]

[1] The Key Laboratory of Intelligent Control and Optimization for Industrial Equipment, Dalian University of Technology, Ministry of Education, Dalian, China
xdzhaohit@gmail.com, wxing@mail.dlut.edu.cn

[2] Department of Mechanical Engineering, Politecnico di Milano, via La Masa 1, 20156 Milan, Italy
hamidreza.karimi@polimi.it

Abstract. Since malicious cyber attacks are common in modern industry, cyber security has become an important issue. Motivated by this, a novel feedback watermark approach is proposed to defend cyber physical systems against replay attacks in this paper. Specifically, we propose a new secure control framework to defend systems against replay attacks, where a feedback channel and a feedback control function are constructed. We address the problem starting from the problem formulation with a linear quadratic Gaussian (LQG) controller, a Kalman filter and a χ^2 detector. We investigate the LQG performance with replay attacks and study the LQG performance loss after adding feedback watermarks. It is also proved that the detector output will converge to a new steady point with a class of feedback function when an attacker launches replay attacks. It is shown that our approach can increase the probability of detecting replay attacks. Finally, we show the validity of feedback watermarks with a numerical example.

Keywords: Cyber physical system · Physical watermark · Replay attack

1 Introduction

Nowadays, due to the development of hardware equipment and telecommunication, more data has been used to guide industrial processes. Networks are more effective to deal with the explosion of data, and hence are frequently used in cyber physical systems (CPSs). However, the openness of networks has caused many drawbacks in security [6]. For example, Stuxnet Worm [13] destroyed the centrifuges in Iran. Such malicious attacks may even be harmful to people's lives.

This work are supported by the National Natural Science Foundation of China (61722302, 61573069), the Liaoning Revitalization Talents Program (XLYC1907140), National Major Science and Technology Project (J2019-V-0010–0105) and the Fundamental Research Funds for the Central Universities (DUT19ZD218).

S. Paiva et al. (Eds.): SmartCity360° 2021, LNICST 442, pp. 702–715, 2022.
https://doi.org/10.1007/978-3-031-06371-8_45

Therefore, cyber security has become a significant issue in industry and detection of malicious attacks is an urgent problem to be solved.

There are two types of malicious attacks widely concerned in the research community. The first one is denial of service (DoS) attack, which enables attackers to jam communication channels. DoS attacks are simple to launch and hence have been adopted by malicious hackers in many cases (see e.g. [7]). However, DoS attack consumes more energy and is not disguised. Signal to interference plus noise ratio (SINR) is frequently used to construct the energy consumption of DoS attacks and researchers have proposed different strategies to defend CPSs under DoS attacks. For example, based on game theory, Li et al. [8] has developed a framework design the transmission strategy, which can be used to defend remote estimation. De Persis et al. [3] has designed secure control in linear systems under DoS attack by appropriate scheduling of transmission time.

False data injection (FDI) attack, the second type of attack, is more destructive since it has the ability to generate malicious signals. However, it is not easy to launch this type of attack since most of the malicious FDI attacks (e.g. [11]) need full knowledge of the system. Detection of false data injection attacks is hard and many people have proposed different methods to detect this type of attack. For instance, Li et al. [7] separated sensors into benign ones and susceptible ones. They designed a fusion algorithm to distinguish malicious sensors.

Replay attack is one of false data injection attacks, which only needs to record sensor measurements and replay them to a fusion center. In other words, it does not need any knowledge about the system. Because of the simplicity of replay attacks, it could be utilized in many scenarios (e.g., the Stuxnet Worm [13]) and even bypassed detectors. Therefore, we concentrated on detection of replay attacks in this paper. To be more specific, we utilize physical watermarks to detect replay attacks.

In order to detect malicious attacks, physical watermarks are small noises added to the control input, which were originally proposed in [12]. Mo et al. [10] have extended the result to general controllers, and they designed a specific detector to detect replay attacks. Weerakkody et al. [15] have used physical watermarks to detect false data injection attacks when some of the controllers are eavesdropped by the attacker. To reduce the performance loss and increase the detection rate, Miao et al. [9] applied a game theoretic approach in the watermark design to detect replay attacks. They developed an algorithm to calculate a suboptimal solution in a finite time case. Huang et al. [5] applied watermarks to guard remote estimation. Satchidanandan et al. [14] have equipped CPSs with Neyman-Pearson detectors and used watermarks to defend systems. Fang et al. [4] added watermarks with a periodic schedule to reduce performance loss.

In this work, a novel feedback watermark approach is proposed to defend CPSs against replay attacks. First, we formulate the whole system and analyze the effects of replay attacks. Furthermore, the system performance is investigated after adding feedback watermarks. Then we introduce feedback watermarks and propose a formula to design the feedback control function. With the above

function, it is proved that feedback watermarks are more efficient. We summarize the main contributions of this paper as follow:

(1) The effects of replay attacks are investigated. To be more specific, we show that the system is stable under replay attack if the system matrix A is Hurwitz.
(2) We propose a novel feedback watermark approach, which can be seen as an extension of the existing works [10]. Moreover, we study the LQG performance loss after adding the watermark. It is shown that the detection rate will increase to a new steady point and our method is more effective than the method in [10].
(3) A design formula is given to design feedback control function, which can be utilized to generate physical watermarks. One can develop specific feedback control function based on our formula.

The rest of this paper is organized as follows: in Sect. II, we formulate a CPS with a Kalman filter, a controller and a χ^2 detector. In Sect. III, we deeply investigate the effects of replay attacks. Section IV proposes the feedback watermark approach and investigates a formula to design feedback control function. In Sect. V gives a numerical example to show the validity and advantages of the proposed methods. Section VI provides some concluding remarks.

Notation: $\mathbb{R}$ represents the set of real numbers. $\mathbb{E}(\cdot)$ is the expectation of a random event. $\mathcal{N}(\mu, \Sigma)$ denotes a Gaussian distribution with mean μ and covariance matrix Σ.

2 Problem Formulation

2.1 System Dynamic and Kalman Filter

Consider the following discrete linear invariant system:

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

where $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}^m$ is the state and the measurement output. $u_k \in \mathbb{R}^l$ is the control input. w_k and v_k is the Gaussian noise with covariance Q and R. It is assumed that w_k and v_j are independent of each other, w_k is independent of $w_j, \forall j \neq k$ and v_k is independent of $v_j, \forall j \neq k$ i.e.,$\mathbb{E}[w_k v_j^T] = 0, \forall k, j$, $\mathbb{E}[w_k w_j^T] = 0, \forall k \neq j$ and $\mathbb{E}[v_k v_j^T] = 0, \forall k \neq j$.

A Kalman filter is used in this paper to estimate the state, it is well known that a Kalman filter can be expressed in an iteration form:

$$\hat{x}_{k|k-1} = A\hat{x}_{k-1|k-1} + Bu_{k-1}, \tag{3}$$
$$P_{k|k-1} = AP_{k-1|k-1}A^T + Q, \tag{4}$$
$$K_k = P_{k|k-1}C^T(CP_{k|k-1}C^T + R)^{-1}, \tag{5}$$
$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k(y_k - C\hat{x}_{k|k-1}), \tag{6}$$
$$P_{k|k} = (I - K_k C)P_{k|k-1}, \tag{7}$$

where $P_{k|k-1} = \mathbb{E}[(x_k - \hat{x}_{k|k-1})(x_k - \hat{x}_{k|k-1})^T]$ and $P_{k|k} = \mathbb{E}[(x_k - \hat{x}_{k|k})(x_k - \hat{x}_{k|k})^T]$. It should be pointed out that a Kalman filter will enter the steady state exponentially fast. Therefore, it is without loss of generality that we assume the Kalman filter run in the steady state at the beginning, i.e.,

$$P_{0|-1} = \bar{P}, K_k = K, \Pi_0 = (I - KC)\bar{P}, \tag{8}$$

$$\hat{x}_{k|k-1} = A\hat{x}_{k|k} + Bu_{k-1}, \tag{9}$$

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K(y_k - C\hat{x}_{k|k-1}), \tag{10}$$

where $\bar{P} \triangleq \lim_{k\to\infty} P_{k|k-1}$, $K \triangleq \bar{P}C^T \left(C\bar{P}C^T + R\right)^{-1}$.

2.2 Linear Quadratic Gaussian (LQG) Optimal Control

The LQG controller in this paper minimizes the infinite time linear quadratic objective function [2] as follows:

$$J = \min \lim_{T\to\infty} \mathbb{E}\{\frac{1}{T}\left[\sum_{k=0}^{T-1}\left(x_k^T W x_k + u_k^T U u_k\right)\right]\}, \tag{11}$$

where W and U are semidefinite matrices. Given the state estimate $\hat{x}_{k|k}$, a fixed gain controller will solve the minimization problem, taking the following form:

$$u_k = u_k^* = -\left(B^T S B + U\right)^{-1} B^T S A \hat{x}_{k|k}, \tag{12}$$

where u_k^* is the optimal control input, and S is the solution of the following Riccati equation [2]:

$$S = A^T S A + W - A^T S B \left(B^T S B + U\right)^{-1} B^T S A. \tag{13}$$

Let us define $L \triangleq -\left(B^T S B + U\right)^{-1} B^T S A$, then

$$u_k^* = L\hat{x}_{k|k}. \tag{14}$$

Using the above fixed gain controller, J becomes a constant dependent on system parameters: [2]:

$$J = \text{trace}(SQ) + \text{trace}\left[\left(A^T S A + W - S\right)(P - KCP)\right]. \tag{15}$$

2.3 χ^2 Detector

A χ^2 detector is a very commonly used detector to detect abnormal behavior in a system. Let us define $z_k = y_k - C\hat{x}_{k|k-1}$. Then, a χ^2 detector takes the following form:

$$g_k = \sum_{i=k-\mathscr{T}+1}^{k} z_i^T \mathscr{P}^{-1} z_i \underset{\mathscr{H}_1}{\overset{\mathscr{H}_0}{\lessgtr}} \eta, \tag{16}$$

where $\mathscr{P} = C\bar{P}C^T + R$, $\mathscr{T}$ is the window size and η is the threshold. $\mathscr{H}_0$ denotes that the system is operating normally, and $\mathscr{H}_1$ denotes that an attacker implements an attack strategy $\mathscr{Z}$. The probability of detection β_k when the system is under attack and the probability of false alarm α_k are defined respectively as

$$\beta_k \triangleq \Pr\left(g_k > \eta \mid \mathscr{H}_1\right), \quad \alpha_k \triangleq \Pr\left(g_k > \eta \mid \mathscr{H}_0\right). \tag{17}$$

In this paper, the window size $\mathscr{T}$ is set to be 1. However, it is trivial to extend our results to the scenario where $\mathscr{T} > 1$ by vector stack.

2.4 Some Lemmas

Before introducing replay attacks, the following lemmas are recalled here to facilitate the understanding of the paper:

Lemma 1. *[1]*

(i). z_k is a Gaussian distributed vector with probability distribution function (PDF) $\sim \mathscr{N}(0, \mathscr{P})$.

(ii). $E[z_k^T z_l] = 0$, $\forall k \neq l$.

Lemma 2. *[10] Define $\mathscr{A} = (A + BL)(I - KC)$. If $\mathscr{A}$ is stable, the detection rate β_k will converge to α_k, i.e.,*

$$\lim_{k\to\infty} \beta_k = \alpha_k. \tag{18}$$

Conversely, if $\mathscr{A}$ is unstable, the detection rate β_k will converge to 1, i.e.,

$$\lim_{k\to\infty} \beta_k = 1. \tag{19}$$

Remark 1. Lemma 1 and Lemma 2 are technical and used in the proof of Theorem 1.

3 Effects of Replay Attacks

In this section, we concentrate on the detector performance provided the attacker is launching a replay attack. Without loss of generality, it is assumed the replay attack begins at time 0. Following [10], we can think of y'_ks (the attacker's records) as outputs of the following virtual system:

$$x'_{k+1} = Ax'_k + Bu'_k + w'_k, \tag{20}$$

$$y'_k = Cx'_k + v'_k, \tag{21}$$

$$\hat{x}'_{k+1|k} = A\hat{x}'_{k|k} + Bu'_k, \tag{22}$$

$$\hat{x}'_{k+1|k+1} = \hat{x}'_{k+1|k} + Kz'_k, \tag{23}$$

$$u'_k = L\hat{x}'_{k|k} + \Delta u'_k, \tag{24}$$

where $\Delta u'_k \sim (0, \mathscr{Q}')$. Define $\hat{x}_{0|-1} - \hat{x}'_{0|-1} = \xi$ and $z_k^a = y'_k - C\hat{x}_{k|k-1}$. It is worth pointing out that z_k^a is the innovation when a replay attack occurs.

Assumption 1. *$\mathscr{A}$ is stable.*

Assumption 2. *x_0 is independent of x'_0, w'_k and v'_k.*

Remark 2. Even though the two assumptions seem very harsh at the first sight, it is reasonable since a replay attack will be detected immediately if $\mathscr{A}$ is unstable. Then, we don't need to design a mechanism to detect replay attacks. Moreover, an attacker is more likely to record a long period measurements. And the replay part signal is almost independent to the real time signal (x_0) due to the system is stable. It should be noticed that $z'_k \triangleq y'_k - Cx'_{k|k-1}$ is independent of x_k since Assumption 2 holds.

Theorem 1. *Under the Assumptions 1 and 2, the system will be stable if A is Hurwitz. When A is Hurwitz, the LQG performance under a long period replay attack without being detected is:*

$$J = \mathrm{trace}(\tilde{W}\bar{P}^a), \tag{25}$$

and $\bar{P}^a \in \mathbb{R}^{2n}$ satisfies the following equation,

$$\bar{P}^a = \tilde{A}\bar{P}^a\tilde{A}^T + \tilde{B}\begin{bmatrix} Q & 0 \\ 0 & \mathscr{P} \end{bmatrix}\tilde{B}^T, \tag{26}$$

where $\tilde{W} = \begin{bmatrix} W + L^TUL & L^TUL \\ L^TUL & L^TUL \end{bmatrix}$, $\tilde{A} = \begin{bmatrix} A + BL & BL \\ 0 & A \end{bmatrix}$ *and* $\tilde{B} = \begin{bmatrix} I & 0 \\ -I & K \end{bmatrix}$.

Proof. Rewrite $\hat{x}_{k+1|k}$ as

$$\hat{x}_{k+1|k} = \mathscr{A}\hat{x}_{k|k-1} + (A + BL)Ky'_k. \tag{27}$$

For the virtual system,

$$\hat{x}'_{k+1|k} = \mathscr{A}\hat{x}'_{k|k-1} + (A + BL)Ky'_k. \tag{28}$$

Then it can be shown that

$$\begin{aligned} z^a_k &= y'_k - C\hat{x}'_{k|k-1} + C\hat{x}'_{k|k-1} - C\hat{x}_{k|k-1} \\ &= z'_k + C(\hat{x}'_{k|k-1} - \hat{x}_{k|k-1}) \\ &= z'_k - C\mathscr{A}^k\xi. \end{aligned} \tag{29}$$

Considering $\mathscr{A}$ is stable, this shows the residue under attack will coverage to the normal residue z'_k, which means that a replay attack can bypass the χ^2 detector.

$$\begin{aligned} \hat{x}_{k+1|k+1} &= \hat{x}_{k+1|k} + Kz_k \\ &= (A + BL)\hat{x}_{k|k} + Kz^a_k \\ &= (A + BL)\hat{x}_{k|k} + Kz'_k - KC\mathscr{A}^k\xi. \end{aligned} \tag{30}$$

The system dynamic (1) combined with (14) can be rewritten as:

$$x_{k+1} = (A + BL)x_k + BLe_k + w_k. \tag{31}$$

Now Let us define $e_k \triangleq \hat{x}_{k|k} - x_k$, $\theta_k = [x_k^T, e_k^T]^T$ and $P_k^a \triangleq \mathbb{E}(\theta_k \theta_k^T)$. Subtracting the above two equations, it can be obtained

$$e_{k+1} = Ae_k + Kz'_k - w_k - KC\mathscr{A}^k\xi, \tag{32}$$

and

$$\theta_{k+1} = \tilde{A}\theta_k + \tilde{B}\begin{bmatrix} w_k \\ z'_k \end{bmatrix} + \tilde{E}\mathscr{A}^k\xi, \tag{33}$$

where $\tilde{A} = \begin{bmatrix} A + BL & BL \\ 0 & A \end{bmatrix}$, $\tilde{B} = \begin{bmatrix} I & 0 \\ -I & K \end{bmatrix}$ and $\tilde{E} = \begin{bmatrix} 0 \\ -KC \end{bmatrix}$. It can be observed that the estimation error will go to infinity when the system parameter A is not Hurwitz and hence the system will be unstable. When A is Hurwitz, one can obtain $\lim_{k\to\infty} P_k^a = \bar{P}^a$, which satisfies the following condition:

$$\bar{P}^a = \tilde{A}\bar{P}^a\tilde{A}^T + \tilde{B}\begin{bmatrix} Q & 0 \\ 0 & \mathscr{P} \end{bmatrix}\tilde{B}^T. \tag{34}$$

It can be obtained that

$$\begin{aligned} J &= \lim_{T\to\infty} \mathbb{E}\{\frac{1}{T}\left[\sum_{k=0}^{T-1}\left(x_k^T W x_k + u_k^T U u_k\right)\right]\} \\ &= \lim_{T\to\infty} \mathbb{E}\{\frac{1}{T}\left[\sum_{k=0}^{T-1}\left(\theta_k^T\begin{bmatrix} W & 0 \\ 0 & 0 \end{bmatrix}\theta_k + \theta_k^T\begin{bmatrix} L^TUL & L^TUL \\ L^TUL & L^TUL \end{bmatrix}\theta_k\right)\right]\} \\ &= \mathrm{trace}(\tilde{W}\bar{P}^a), \end{aligned} \tag{35}$$

where $\tilde{W} = \begin{bmatrix} W + L^TUL & L^TUL \\ L^TUL & L^TUL \end{bmatrix}$. This ends the proof. ■

Remark 3. Since a replay attack is usually launched for a long time [13] and $\mathscr{A}$ is a Hurwitz matrix, it is reasonable to omit the last term of (30). Moreover, if an attacker has the ability to eavesdrop the sensor measurements, the time to begin replay attacks will be chosen carefully to avoid being detected, meaning that y_0 and y'_0 will be chosen to be approximately the same and hence ξ will be small. Therefore, a replay attack cannot be detected easily when $\mathscr{A}$ is stable

4 Detection of Replay Attacks

In order to defend against replay attacks. Mo et al. [12] proposed a physical watermark approach, where physical watermarks denote small noises in the control inputs. To be more specific, the watermark signals are only known to the system and the attacker has no information about it. In order to efficiently

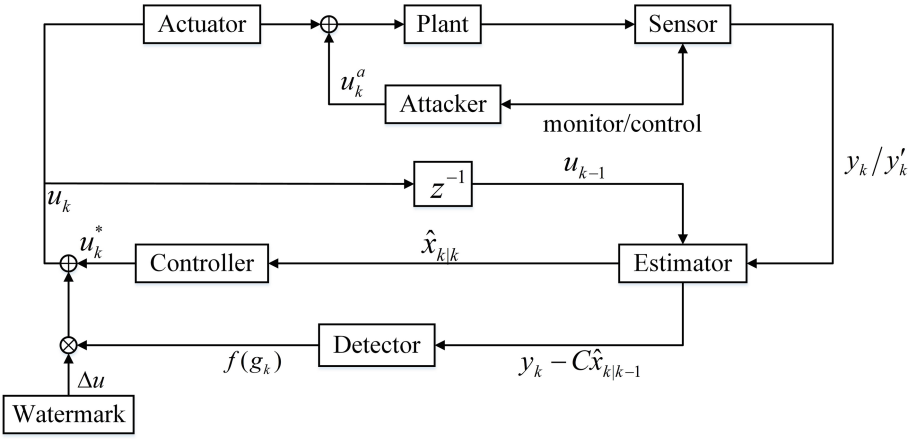

Fig. 1. System architecture

detect malicious attacks, we propose a feedback watermark approach, where we connect the physical watermark with the detector output (See Fig. 4). Let $\Delta u_k^* \sim \mathcal{N}(0, \mathcal{Q})$ denote the ordinary watermark which is independent of x_0, w_k and v_k for all k, our feedback watermark is given by:

$$\Delta u_k = f(g_k)\Delta u_k^*, \tag{36}$$

where $f(\cdot) := \mathbb{R}^+ \to [0, \delta]$ is a measurable function to control the watermark signal. Therefore, we call this function as a feedback control function. The reason why we adopt an upper bound δ here is to avoid an unbearable physical watermark. Otherwise, a small malicious signal may lead a serious consequence in the control system. In other words, we would like to adopt a mechanism to increase the covariance of watermarks with a relatively small impact on the system when the system is under attack. Furthermore, if $f(g_k) = 1$, we may obtain the ordinary watermark as in [10], meaning that our approach is an extension to the existing ones.

In the next part, we want to explore the effects of the feedback control functions $f(\cdot)$. $f(\cdot)$ can be designed according to Theorem 2. First, let us explore the LQG performance loss after adding physical watermarks.

Theorem 2. *When there is no attack, the LQG performance J' after adding feedback physical watermark is given by*

$$J' = J + \Delta J, \tag{37}$$

where $\Delta J = \tilde{h}\,\mathrm{trace}\left(\left(B^T S B + U\right)\mathcal{Q}\right)$, *with* $\tilde{h} = \mathbb{E}[f(g_0)^2]$.

Proof. The proof is similar to the proof of Theorem 1 in [12] and hence is omitted here. ■

The next theorem shows that the covariance of Δu_k is larger when the system is under attack, i.e., $\mathbb{E}[f(g_k)^2] \geq \mathbb{E}[f(g'_k)^2]$. Therefore, our approach can be used to defend systems against replay attacks.

Theorem 3. *Let $h(x) \triangleq f(x)^2$. If $h(x)$ is a monotone increasing truncation function, then the following statements hold:*
(i) $\mathbb{E}[h(g_k)]$ will converge to a steady point $\bar{h}$,
(ii) $\lim_{k\to\infty} \mathbb{E}(g_k) = m + \text{trace}\left(C^T \mathscr{P}^{-1} C \bar{\Sigma}\right)$, where $\bar{\Sigma} = \mathscr{U} + \mathscr{W}$, $\mathscr{U}$ satisfies $\mathscr{U} - \bar{h} B \mathcal{Q} B^T = \mathscr{A} \mathscr{U} \mathscr{A}^T$ and $\mathscr{W}$ satisfies $\mathscr{W} - \bar{h} B \mathcal{Q} B^T = \mathscr{A} \mathscr{W} \mathscr{A}^T$.

Proof. The system dynamic now becomes:

$$x_{k+1} = Ax_k + Bu_k + B\Delta u_k + w_k, \tag{38}$$

and thus z_k^a can be recalculated as

$$\begin{aligned} z_k^a &= y'_k - C\hat{x}'_{k|k-1} + C\hat{x}'_{k|k-1} - C\hat{x}_{k|k-1} \\ &= z'_k + C\mathscr{A}(\hat{x}'_{k-1|k-2} - \hat{x}_{k-1|k-2}) - CB(\Delta u_{k-1} - \Delta u'_{k-1}) \\ &= z'_k - C\mathscr{A}^k \xi - C\sum_{i=0}^{k-1} \mathscr{A}^{k-i-1} B(\Delta u_i - \Delta u'_i) \\ &= z'_k - C\mathscr{A}^k \xi - C\sum_{i=0}^{k-1} \mathscr{A}^{k-i-1} B(f(g_k^a)\Delta u_i^* - f(g'_k)\Delta u_i^{*\prime}). \end{aligned} \tag{39}$$

Then, it can be calculated that

$$\begin{aligned} \mathbb{E}(z_k^a z_k^{aT}) = &\mathbb{E}\{z'_k z_k'^T + C\mathscr{A}^k \xi\xi^T (\mathscr{A}^k)^T C^T \\ &+ C\sum_{i=0}^{k-1} h(g_i^a)\mathscr{A}^{k-i-1} B\Delta u_i \Delta u_i^T B^T (\mathscr{A}^{k-i-1})^T C^T \\ &+ C\sum_{i=0}^{k-1} h(g'_i)\mathscr{A}^{k-i-1} B\Delta u_i^* \Delta u_i^{*T} B^T (\mathscr{A}^{k-i-1})^T C^T\} \\ = &\mathscr{P} + C\mathscr{A}^k \xi\xi^T (\mathscr{A}^k)^T C^T + C\sum_{i=0}^{k-1} \mathbb{E}[h(g_i^a)]\mathscr{A}^{k-i-1} B\mathcal{Q}B^T (\mathscr{A}^{k-i-1})^T C^T \\ &+ C\sum_{i=0}^{k-1} \mathbb{E}[h(g'_i)]\mathscr{A}^{k-i-1} B\mathcal{Q}B^T (\mathscr{A}^{k-i-1})^T C^T\}, \end{aligned} \tag{40}$$

where the superscript a helps to clarify vectors under attack. Since $C\mathscr{A}^k \xi\xi^T (\mathscr{A}^k)^T C^T$ will be close to zero arbitrarily when k is large enough, this term is omitted in the following proof. Now (40) becomes

$$\begin{aligned} \mathbb{E}(z_k^a z_k^{aT}) = &\mathscr{P} + C\sum_{i=0}^{k-1} \mathbb{E}[h(g_i^a)]\mathscr{A}^{k-i-1} B\mathcal{Q}B^T (\mathscr{A}^{k-i-1})^T C^T \\ &+ C\sum_{i=0}^{k-1} \mathbb{E}[h(g'_i)]\mathscr{A}^{k-i-1} B\mathcal{Q}B^T (\mathscr{A}^{k-i-1})^T C^T\}. \end{aligned} \tag{41}$$

First, we show that $\mathbb{E}[h(g_1^a)] \geq \mathbb{E}[h(g_0^a)]$. Define $\Sigma_k \triangleq \mathbb{E}(z_k^a z_k^{aT})$, it is trivial to prove $\Sigma_1 \geq \Sigma_0$. Also, let $v_0^a = \mathscr{P}^{\frac{1}{2}} z_0^a \sim \mathscr{N}(0, \mathscr{P}^{\frac{1}{2}} \Sigma_0 \mathscr{P}^{\frac{1}{2}})$ and $v_1^a = \mathscr{P}^{\frac{1}{2}} z_1^a \sim \mathscr{N}(0, \mathscr{P}^{\frac{1}{2}} \Sigma_1 \mathscr{P}^{\frac{1}{2}})$. Suppose z_k^a is a scalar, i.e., $m = 1$, it can be shown that (recall that the window size $\mathscr{T} = 1$)

$$\begin{aligned}
\mathbb{E}[h(g_1^a)] - \mathbb{E}[h(g_0^a)] =& \mathbb{E}[h\left(z_1^{aT} \mathscr{P}^{-1} z_1^a\right)] - \mathbb{E}[h\left(z_0^{aT} \mathscr{P}^{-1} z_0^a\right)] \\
=& \int_{\mathbb{R}} h(v_1^{a2}) p_1(v_1^a) dv_1^a - \int_{\mathbb{R}} h(v_0^{a2}) p_0(v_0^a) dv_0^a \\
=& \int_{\mathbb{R}} h(v^2) \left(p_1(v) - p_0(v)\right) dv \\
=& \int_{\mathbb{V}} h(v^2) \left(p_1(v) - p_0(v)\right) dv + \int_{\mathbb{R}\setminus\mathbb{V}} h(v^2) \left(p_1(v) - p_0(v)\right) dv \\
\geq& \int_{\mathbb{V}} h(\bar{v}^2) \left(p_1(v) - p_0(v)\right) dv + \int_{\mathbb{R}\setminus\mathbb{V}} h(\bar{v}^2) \left(p_1(v) - p_0(v)\right) dv \\
=& 0, \qquad (42)
\end{aligned}$$

where $\mathbb{V} = \{v | p_1(v) \leq p_0(v)\}$ and $\bar{v}$ satisfies $p_1(\bar{v}) = p_0(\bar{v})$. Since the density function of a Gaussian distribution is symmetrical and h is a monotone increasing function, the inequality holds. When z_k^a is a vector, the additional step in the proof is to diagonalize the covariance matrix first and then the following proof is similar. Hence, we omit the detailed proof for the general case here.

Now, it is trivial to see that $\Sigma_2 \geq \Sigma_1$ and $\mathbb{E}[h(g_2^a)] \geq \mathbb{E}[h(g_1^a)]$. Following the same procedure, it can be shown that $\mathbb{E}[h(g_k + 1^a)] \geq \mathbb{E}[h(g_k^a)]$. Combined this with $\mathbb{E}[h(g_k^a)] \leq \delta^2$, it can be concluded that $\lim_{k\to} \mathbb{E}[h(g_k^a)] = \bar{h}$, which proves (i). To prove (ii), it can be shown from (i) and the last two terms of (41) will converge to the steady state given by the Lyapunov equations that

$$\lim_{k\to\infty} \Sigma_k = \bar{\Sigma}. \qquad (43)$$

It can be calculated that

$$\begin{aligned}
\lim_{k\to\infty} \mathbb{E}(g_k) &= \lim_{k\to\infty} \mathbb{E}(z_k^{aT} \mathscr{P}^{-1} z_k^a) \\
&= \lim_{k\to\infty} \operatorname{trace}\left(\mathscr{P}^{-1} \Sigma_k\right) \\
&= \operatorname{trace}\left(\mathscr{P}^{-1} (\mathscr{P} + C\bar{\Sigma} C^T)\right) \\
&= m + \operatorname{trace}\left(C^T \mathscr{P}^{-1} C \bar{\Sigma}\right), \qquad (44)
\end{aligned}$$

which finishes the proof. ■

Remark 4. It should be emphasized that h is a monotone increasing function since we would like to generate a larger physical watermark when the system is in anomaly. It is also worth pointing out $\bar{h} \geq \tilde{h}$ by $\mathbb{E}[h(g_k^a)] \geq \mathbb{E}[h(g_0^a)]$, meaning that our approach is more effective when the ΔJ is the same. This shows that our method will have a higher detection rate compared with that in [10] when the performance loss is the same. Furthermore, the detector proposed in [10] can

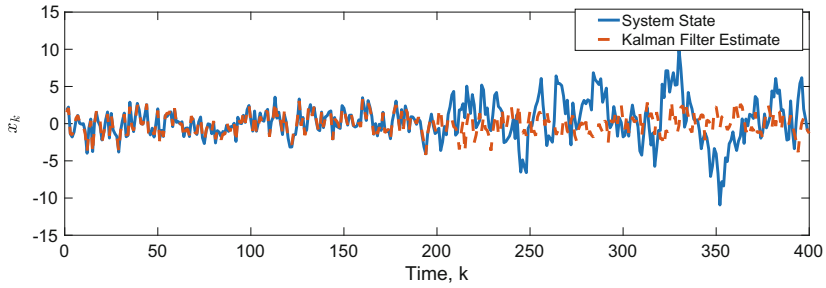

Fig. 2. System state and Kalman Filter estimate under replay attacks.

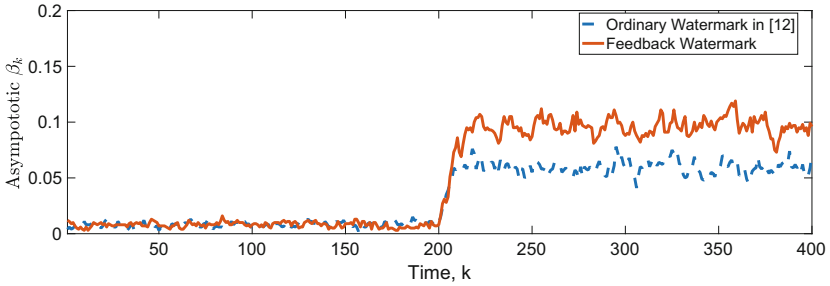

Fig. 3. Comparisons between feedback watermark and ordinary watermark in [10].

also be utilized in this paper. Actually, any detector using statistical properties of the innovation z_ks can be used in this paper. We adopt χ^2 detector in this paper for the brevity of explanations.

5 Numerical Example

In this section, we provide a numerical example to show the effectiveness of our approach. The system parameters are as follow: $A = 0.8$, $B = 1$, $C = 1$, $Q = 1$, $R = 0.1$, $W = 1$, $U = 1$, $\mathscr{T} = 10$ and the feedback function is chosen as $f(x) = \sqrt{x}$, which means that the square root of the output of χ^2 detector is directly multiplied with the ordinary Δu_k^*. The upper bound δ is chosen as 5. Under this condition, $\tilde{h} = m\mathscr{T} = 10$ is the standard expectation of a χ^2 detector. ΔJ with feedback watermarks can be calculated by Theorem 2 as $\Delta J = 23.7\mathcal{Q}$.

It can be seen from Fig. 2, the estimate deviates from the system states when the replay attack begins at time step 201. Therefore, the controller can not have a satisfactory LQG performance.

In order to compare with the existing watermark approach in [10], we make ΔJ the same in the two methods. Thus, we set $\mathcal{Q} = 0.02$ in our method and $\mathcal{Q} = 0.2$ in the ordinary watermark. Moreover, the false alarm rate is set to be 0.001. The attacker records the measurements from time 1 to time 200 and

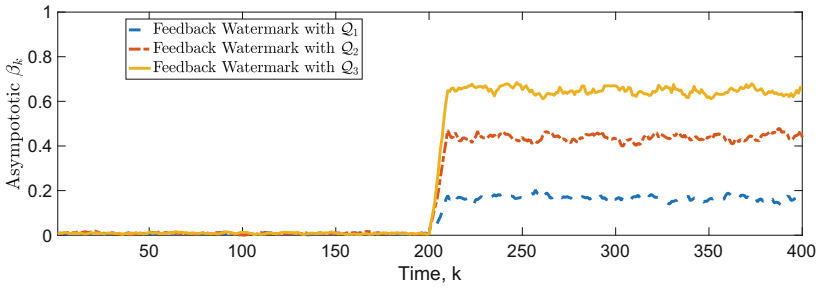

Fig. 4. Comparisons between feedback watermark with different $\mathcal{Q}$s.

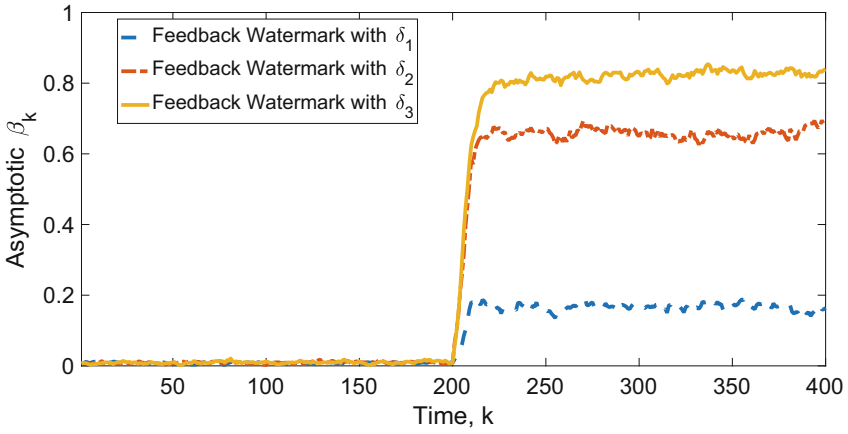

Fig. 5. Comparisons between feedback watermark with different δs.

replay them between time 201 and time 400. The results are the average of 1000 times experiments.

From Fig. 3, one can see that feedback watermark is more effective than the ordinary one. We wish to investigate the detection rates with different $\mathcal{Q}$s. To do this, we set $\delta = 2$ and $\mathcal{Q}_1 = 0.1$, $\mathcal{Q}_1 = 0.2$, $\mathcal{Q}_1 = 0.3$. From Fig. 4, one can find that a larger $\mathcal{Q}$ can lead a larger detection rate.

To better illustrate the effect of the upper bound δ, we set $\delta_1 = 2$, $\delta_2 = 4$, $\delta_3 = 6$. In this case, $\mathcal{Q} = 0.1$. It is important to recall here that a larger δ means the system can endure a larger noise. It is immediately seen from Fig. 5, a larger δ is beneficial to the detection rates. However, the choice of a larger δ increases the uncertainty of the whole system.

We also give a diagram Fig. 6 demonstrating the detection rates with different parameters. A system designer can choose the feedback control function according to this diagram.

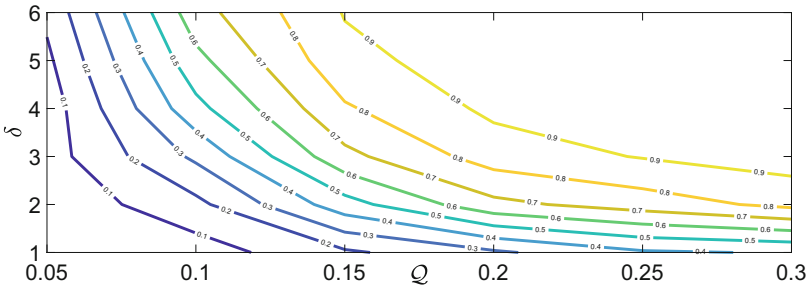

Fig. 6. Detection rates with different $\mathcal{Q}$s and δs.

Finally, it is worth pointing out here that our feedback approach can be utilized to generate dependent watermark in [11]. Moreover, other works mentioned in the introduction mainly focus on other scenarios. Hence, we omit the other comparisons here.

6 Conclusions

In this paper, a feedback watermark approach has been designed to detect malicious replay attacks. First, the effects of replay attacks have been analyzed. Then, we discussed the LQG performance loss after adding feedback watermark. It has been proved that the detector output converge to a steady point when a class of feedback function is utilized, meaning that our approach is more effective than the ordinary watermark. Finally, we have shown the validity of feedback watermarks with a numerical example. However, since the feedback function can be designed in different forms, we will explore more efficient feedback functions in future work.

References

1. Anderson, B., Moore, J.B.: Optimal filtering. Prentice-Hall, Englewood Cliffs (1979)
2. Berstekas, D.P.: Dynamic programming and optimal control. Athena scientific (1995)
3. De Persis, C., Tesi, P.: Input-to-state stabilizing control under denial-of-service. IEEE Trans. Autom. Control **60**(11), 2930–2944 (2015)
4. Fang, C., Qi, Y., Cheng, P., Zheng, W.X.: Optimal periodic watermarking schedule for replay attack detection in cyber-physical systems. Automatica **112**, 108698 (2020). https://doi.org/10.1016/j.automatica.2019.108698
5. Huang, J., Ho, D.W., Li, F., Yang, W., Tang, Y.: Secure remote state estimation against linear man-in-the-middle attacks using watermarking. Automatica **121**, 109182 (2020). https://doi.org/10.1016/j.automatica.2020.109182
6. Lee, E.A.: Cyber physical systems: design challenges. In: 2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing, ISORC (2008)

7. Li, Y., Shi, L., Chen, T.: Detection against linear deception attacks on multi-sensor remote state estimation. IEEE Trans. Control Network Syst. **5**(3), 846–856 (2018). https://doi.org/10.1109/TCNS.2017.2648508
8. Li, Y., Shi, L., Cheng, P., Chen, J., Quevedo, D.E.: Jamming attacks on remote state estimation in cyber-physical systems: a game-theoretic approach. IEEE Trans. Autom. Control (2015). https://doi.org/10.1109/TAC.2015.2461851
9. Miao, F., Zhu, Q., Pajic, M., Pappas, G.J.: A hybrid stochastic game for secure control of cyber-physical systems. Automatica **93**, 55–63 (2018). https://doi.org/10.1016/j.automatica.2018.03.012, http://www.sciencedirect.com/science/article/pii/S0005109818300992
10. Mo, Y., Chabukswar, R., Sinopoli, B.: Detecting integrity attacks on SCADA systems. IEEE Trans. Control Syst. Technol. **22**(4), 1396–1407 (2014)
11. Mo, Y., Weerakkody, S., Sinopoli, B.: Physical authentication of control systems: designing watermarked control inputs to detect counterfeit sensor outputs. IEEE Control Syst. Mag. **35**(1), 93–109 (2015)
12. Mo, Y., Sinopoli, B.: Secure control against replay attacks. In: Conference on Communication, Control, and Computing (2009)
13. Langner, R.: To kill a centrifuge a technical analysis of what stuxnet 's creators tried to achieve. Technical report, November 2013. www.langner.com/en/wp-content/uploads/2013/11/To-kill-a-centrifuge.pdf
14. Satchidanandan, B., Kumar, P.R.: Dynamic watermarking: active defense of networked cyber-physical systems. Proc. IEEE **105**(2), 219–240 (2017)
15. Weerakkody, S., Ozel, O., Sinopoli, B.: A bernoulli-gaussian physical watermark for detecting integrity attacks in control systems. In: 2017 55th Annual Allerton Conference on Communication, Control, and Computing, Allerton, pp. 966–973 (2017)