Sequential detection of Replay attacks

Arunava Naha[1], André Teixeira[2], Anders Ahlén[1] and Subhrakanti Dey[3]

***Abstract*—One of the most studied forms of attacks on the cyber-physical systems is the replay attack. The statistical similarities of the replayed signal and the true observations make the replay attack difficult to detect. In this paper, we address the problem of replay attack detection by adding watermarking to the control inputs and then perform resilient detection using cumulative sum (CUSUM) test on the joint statistics of the innovation signal and the watermarking signal, whereas existing work considers only the marginal distribution of the innovation signal. We derive the expression of the Kullback-Liebler divergence (KLD) between the two joint distributions before and after the replay attack, which is, asymptotically, inversely proportional to the detection delay. We perform a structural analysis of the derived KLD expression and suggest a technique to improve the KLD for the systems with relative degree greater than one. A scheme to find the optimal watermarking signal variance for a fixed increase in the control cost to maximize the KLD under the CUSUM test is presented. We provide various numerical simulation results to support our theory. The proposed method is also compared with a state-of-the-art method based on the Neyman-Pearson detector, illustrating the smaller detection delay of the proposed sequential detector.**

***Index Terms*—Replay attack, sequential detection, CUSUM test, Networked control system.**

I. Introduction

Nowadays, large-scale cyber-physical systems (CPS) are getting deployed for intelligent transportation systems, manufacturing industries, smart grids, etc. [1]. Along with their immense advantages, there are also growing concerns about the safety and security of such systems. Attacks on the CPS can be a serious threat to the sensitive user data security, availability and reliability of critical resources, user's physical safety, and monetary loss [1]. Various techniques, such as data encryption, authentication, firewall, cryptography, digital watermarking, etc. are normally deployed to protect the cyber-layer of the CPS. Such protection schemes may not be adequate to protect the CPS from attacks on the physical layers as realised from different past incidents, such as the famous Stuxnet attack [2]. In the Stuxnet attack, the malware issued harmful control inputs to increase the pressure of the centrifuges in a uranium enrichment plant in Iran [3]. It also replaced the true measurements with previously recorded observations to remain stealthy. An attacker can launch a replay attack without detailed knowledge about the system parameters and control logic. The attacker can hijack a sensor node and record the observation for some time, and then replay it back by replacing the true measurements at some later point in time. The attacker can alter the system in some harmful way, for example, the attacker can add harmful exogenous inputs and may remain stealthy during the replay attack.

A widely applied technique for the replay attack detection is to add a watermarking signal to the control inputs, and then perform various statistical tests using the observations or the innovation signal from the Kalman estimator [3]–[5]. In one approach, χ^2 statistics generated using the innovation signal is compared with some threshold for attack detection [3], [6], [7]. In another approach, test statistics are built using the observation data to perform a threshold check [4] or the Neyman-Pearson (NP) test [8]. Addition of watermarking increases the probability of detection, but at the same time, it increases the control cost [3]. In [3], [9], optimal watermarking signals are designed which maximizes the attack detectability for a fixed increase in the control cost. In a different approach, the watermarking signal is also added or multiplied with the observations before the transmission. At the receiver, the authenticity of the observations are first checked, and then the watermarking signal is filtered out before feeding the observations to the estimator or controller [10]–[13]. The watermarking signals for the observations could be of different types, such as sinusoidal [10], multiplicative to the observations [11], time-varying sinusoidal [12], random noise [13], etc. Since the added watermarking signal is removed before the observations are fed to the controller, such methods do not increase the control cost. However, if the attacker can access the signal before the addition of the watermarking, then these methods may fail. In [14], the authors design a periodic watermarking scheme for the replay attack detection, which reduces the cost of adding the watermarking to the control inputs during all the time before the attack. On the other hand, transfer entropy-based causality countermeasures are introduced in [15] for four different types of attack detections, including replay attacks, in CPS without using the physical watermarking. Even though most of the methods found in the literature studied the problem of replay attack detection for linear time-invariant (LTI) systems, a detection scheme is reported in [16] for time-varying systems by adding time-varying dynamic watermarking. There are few other methods found in the literature which do not use the watermarking for the replay attack detection. In [17], timestamps are added to the data, and in [18], a nonlinear element is inserted in the control loop for the replay attack detection. A set membership-based approach is followed in [19]. In addition to the research on replay attack detections, researchers have also studied the closed-loop stability of nonlinear systems under attack [20], the conditions on the watermarking to guarantee detection of the replay attack on DC microgrids [21], and the state estimation problem when the system is under attack [22], [23].

Detection of an attack as early as possible is of immense

*This work is supported by The Swedish Research Council (VR) under grants 2017-04053 and 2018-04396, and by the Swedish Foundation for Strategic Research.

[1]Arunava Naha and Anders Ahlén are with the Department of Electrical Engineering, Uppsala University, 75103 Uppsala, Sweden arunava.naha@angstrom.uu.se and Anders.Ahlen@angstrom.uu.se

[2] André Teixeira is with the Department of Information Technology, Uppsala University, 75105 Uppsala, Sweden andre.teixeira @ it.uu.se

[3]Subhrakanti Dey is with the Department of Electronic Engineering, Hamilton Institute, National University of Ireland, Maynooth, Ireland. He is also with the Department of Electrical Engineering, Uppsala University, 75103 Uppsala, Sweden Subhra.Dey@signal.uu.se

importance for the CPS to reduce the magnitude of the damage. Most of the detection mechanisms reported in the literature do not address the issue of detection delay of attacks explicitly. Moreover, some of the reported methods do batch processing which makes the detection delay dependent on the choice of the window size. In addition to that, since the processes are expected to run for a very long time before the attack takes place, the average run length (ARL) between two false alarms is a better metric to use compared to the false alarm rate (FAR) [24]. In contrast to methods based on batch processing, sequential detection schemes recursively process the data online. Furthermore, a typical class of sequential detection schemes, the cumulative sum (CUSUM) test, directly considers ARL when designing the detection threshold. Therefore, we have applied a CUSUM test [25], [26] using the joint distributions of the innovation signal and the watermarking signal before and after the replay attack, where the added watermarking signal is independent and identically distributed (iid). The detection delay is asymptotically inversely proportional to the Kullback–Leibler divergence (KLD) measure between the joint distributions before and after the attack, [25], [26].

In a related paper [27], we have extensively studied two CUSUM tests, optimal CUSUM and sub-optimal CUSUM for the quickest detection of data deception attacks. In the data deception attack model considered in [27], the attacker replaces the true observation with a Gaussian stochastic process. In the current paper, we reformulated the problem and extend the work in [27] to detect replay attacks on the CPS. Our main contributions are as follows.

(i) We have reformulated the sub-optimal CUSUM test from [27] for the replay attack detection and investigated its performance in terms of the average detection delay (ADD) and the increase in the control cost for a fixed upper bound on the ARL.

(ii) We have derived an expression for the KLD between the joint distributions before and after the replay attack.

(iii) We have studied the effect of relative degree on the KLD for the replay attack case and exhibited a way to improve the KLD for systems with a relative degree greater than one.

(iv) A technique to optimize the watermarking signal variance, which maximizes the KLD for a fixed upper bound on the increase in the control cost is also proposed.

The paper is organized as follows. The system model and the attack model are described in Section II. Section III provides the replay attack detection scheme, the KLD expression for the replay attack, and the technique to improve the KLD for systems with a relative-degree greater than one. A technique for optimizing the watermarking signal variance is also discussed in Section III. Section IV provides associated numerical results and Section V concludes the paper.

II. System and Attack Model

This section discusses the system models during normal operations and under replay attacks used in this paper.

A. System Model during Normal Operations

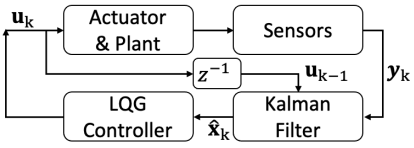

Figure 1: Schematic diagram of the system during normal operation.

Figure 1 shows the schematic diagram of the network control system (NCS) under normal operation employed for this paper. The system is modelled as,

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{w}_k, \tag{1}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{v}_k. \tag{2}$$

Here $\mathbf{x}_k \in \mathbb{R}^n$, $\mathbf{u}_k \in \mathbb{R}^p$, and $\mathbf{y}_k \in \mathbb{R}^m$ are the state, input vector, and output vector at the k-th time instant, respectively. $\mathbf{w}_k \in \mathbb{R}^n \sim \mathcal{N}(0, \mathbf{Q})$ and $\mathbf{v}_k \in \mathbb{R}^m \sim \mathcal{N}(0, \mathbf{R})$ are the iid process noise and observation noise, respectively. $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$, $\mathbf{C} \in \mathbb{R}^{m\times n}$, $\mathbf{Q} \in \mathbb{R}^{n\times n}$, and $\mathbf{R} \in \mathbb{R}^{m\times m}$. The noise vectors $\mathbf{v}_k$ and $\mathbf{w}_k$ are mutually independent, and both are independent of the initial state vector, $\mathbf{x}_0$. We assume the system is stabilizable and detectable. We also assume that the system has been operational from $k = -\infty$, thus the system is at a steady-state from $k \geq 0$. The states are estimated using a Kalman filer as follows,

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{A}\hat{\mathbf{x}}_{k-1|k-1} + \mathbf{B}\mathbf{u}_{k-1}, \tag{3}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}\gamma_k, \tag{4}$$

where $\hat{\mathbf{x}}_{k|k-1} = E[\mathbf{x}_k|\Psi_{k-1}]$ and $\hat{\mathbf{x}}_{k|k} = E[\mathbf{x}_k|\Psi_k]$ are the predicted and filtered state estimates, respectively. $E[\cdot]$ denotes the expected value, and Ψ_k is the set of all measurements up to time k. The innovation γ_k and the steady state Kalman filter gain $\mathbf{K}$ are given by

$$\gamma_k = \mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_{k|k-1}, \tag{5}$$

$$\mathbf{K} = \mathbf{P}\mathbf{C}^T\left(\mathbf{C}\mathbf{P}\mathbf{C}^T + \mathbf{R}\right)^{-1}, \tag{6}$$

where $\mathbf{P} = E\left[(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1})(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1})^T\right]$ is the steady-state error covariance matrix obtained from the following algebraic Riccati equation,

$$\mathbf{P} = \mathbf{A}\mathbf{P}\mathbf{A}^T + \mathbf{Q} - \mathbf{A}\mathbf{P}\mathbf{C}^T\left(\mathbf{C}\mathbf{P}\mathbf{C}^T + \mathbf{R}\right)^{-1}\mathbf{C}\mathbf{P}\mathbf{A}^T. \tag{7}$$

The control input $\mathbf{u}_k$ is generated by minimizing the infinite horizon LQG cost, [3], where the optimal $\mathbf{u}_k^*$ is

$$\mathbf{u}_k^* = \mathbf{L}\hat{\mathbf{x}}_{k|k}, \tag{8}$$

$$\mathbf{L} = -\left(\mathbf{B}^T\mathbf{S}\mathbf{B} + \mathbf{U}\right)^{-1}\mathbf{B}^T\mathbf{S}\mathbf{A}, \tag{9}$$

where $\mathbf{S}$ is the solution to the following algebraic Riccati equation,

$$\mathbf{S} = \mathbf{A}^T\mathbf{S}\mathbf{A} + \mathbf{W} - \mathbf{A}^T\mathbf{S}\mathbf{B}\left(\mathbf{B}^T\mathbf{S}\mathbf{B} + \mathbf{U}\right)^{-1}\mathbf{B}^T\mathbf{S}\mathbf{A}. \tag{10}$$

Here, $\mathbf{W} \in \mathbb{R}^{n\times n}$ and $\mathbf{U} \in \mathbb{R}^{p\times p}$ are positive definite diagonal weight matrices, which are used to formulate the LQG cost [3].

B. Attack Model

As illustrated in the schematic diagram of the system under replay attack in Fig. 2, the attacker has access to the true observations from the system and can record the true observations for a finite but sufficiently long time interval in a buffer to use them at a later point in time. Under the replay attack, the true observation $\mathbf{y}_k$ are replaced by the delayed version of the observation, *i.e.*, $\mathbf{y}_{k-k_0}$, where k_0 represents the delay. The older observation $\mathbf{y}_{k-k_0}$ was recorded by the attacker k_0 time units ago during normal system operation for a finite time interval. Under practical situations, the value of k_0 may vary depending on when the compromised sensors become accessible to the attacker. Since we have assumed that the

system has been operational from $k=-\infty$, and it is at a steady-state from the beginning of the start of observation, *i.e.*, $k \geq 0$, the exact value of k_0 does not have any effect on the CUSUM statistics (14) and the KLD (33). The attacker does not need to have any knowledge about the system parameters or the control logic to launch the replay attack. We assume the attack to start at time $k=\nu$, which is deterministic but unknown. The innovation signal $\widetilde{\gamma}_k$ during the replay attack becomes as follows

$$\widetilde{\gamma}_k = \mathbf{y}_{k-k_0} - \mathbf{C}\hat{\mathbf{x}}^F_{k|k-1}, \tag{11}$$

where $\hat{\mathbf{x}}^F$ denotes the estimated state when the system is under attack.

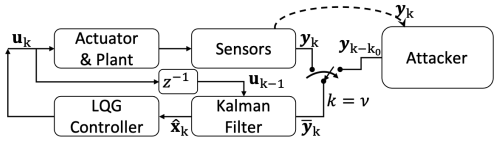

Figure 2: Schematic diagram of the system under replay attack.

III. Replay Attack Detection

This section discusses the replay attack detection scheme, the derivations of the KLD and some other relevant quantities. It also shows a way to optimize the watermarking signal variance, and how to improve the KLD for systems with relative degree greater than one.

A. Detection Scheme

To detect the replay attack, we perform the following two main steps.

Step-1: Add an iid watermarking signal $\mathbf{e}_k \sim \mathcal{N}(0, \mathbf{\Sigma}_e)$ to the optimal LQG control input $\mathbf{u}^*_k$ yielding

$$\mathbf{u}_k = \mathbf{u}^*_k + \mathbf{e}_k. \tag{12}$$

Step-2: Perform the sub-optimal CUSUM test using the joint distributions $f(\gamma_k, \mathbf{e}_{k-1})$ and $\widetilde{f}(\widetilde{\gamma}_k, \mathbf{e}_{k-1})$ of the innovation signal and the watermarking signal, before and after the attack, respectively, as stated in Proposition 1. We compare the CUSUM statistics with a threshold to select from the following two hypotheses,

- H_0: No attack. The estimator receives the true observation $\mathbf{y}_k$
- H_1: Attack. The estimator receives the delayed observation $\mathbf{y}_{k-k_0}$.

Remark 1. *In comparison with the sub-optimal CUSUM test used in this paper for replay attack detection, see (14), the test statistics for the optimal CUSUM test would be [27],*

$$gd_k = \max\left(0, gd_{k-1} + \log \frac{\widetilde{f}\left(\bar{\gamma}_k, \mathbf{e}_{k-1} | \{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-2}\right)}{f\left(\bar{\gamma}_k, \mathbf{e}_{k-1}\right)}\right), \tag{13}$$

where $\{\bar{\gamma}\}_1^{k-1} = \{\gamma_i : 1 \leq i < \nu\} \cup \{\widetilde{\gamma}_i : \nu \leq i \leq k-1\}$ *and* $\{\mathbf{e}\}_1^{k-2} = \{\mathbf{e}_i : 1 \leq i \leq k-2\}$. *The innovation signal becomes dependent on it's previous samples and watermarking signal after the attack. The derivations of the closed-form expressions for the dependent mean and variance of the innovation signal become extremely complex for the replay attack scenario. On the other hand, if we ignore the dependency of $\widetilde{\gamma}_k$ on it's past values, then the CUSUM statistics becomes simpler as stated in Corollary 1.1 of [27]. However, under such an assumption, the CUSUM test will not remain optimal.*

Proposition 1. *The sub-optimal CUSUM test statistics g_k to detect replay attacks is evaluated using Corollary 1.1 from [27] as follows,*

$$g_k = \max\left(0, g_{k-1} + \log \frac{\widetilde{f}\left(\bar{\gamma}_k, \mathbf{e}_{k-1}\right)}{f\left(\bar{\gamma}_k, \mathbf{e}_{k-1}\right)}\right), \tag{14}$$

where $\bar{\gamma}_k = \gamma_k$ before attack, and $\bar{\gamma}_k = \widetilde{\gamma}_k$ after attack. $f(\cdot)$ and $\widetilde{f}(\cdot)$ denote the probability density functions (PDF) before and after the attack, respectively. The decision of attack or no attack is made based on the following,

$$\begin{aligned} H_0 &: \text{ Selected, when } g_k < \log(ARL_h) \\ H_1 &: \text{ Selected, when } g_k \geq \log(ARL_h). \end{aligned} \tag{15}$$

Here, ARL_h is the user selected threshold on ARL, $ARL \geq ARL_h$.

To implement the proposed CUSUM test, the log-likelihood ratio in (14) needs to be computed. In the following, we characterize the probability density functions and derive their expressions. The innovation signal γ_k during the normal operation of the system is uncorrelated to the watermarking signal, see (16). However, on the contrary, the innovation signal $\widetilde{\gamma}_k$ under the replay attack becomes dependent on the watermarking signal, see (17).

$$\begin{aligned} \gamma_k &= \mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_{k|k-1} \\ &= \mathbf{CA}\left(\mathbf{x}_{k-1} - \hat{\mathbf{x}}_{k-1|k-1}\right) + \mathbf{C}\mathbf{w}_{k-1} + \mathbf{v}_k, \end{aligned} \tag{16}$$

$$\begin{aligned} \widetilde{\gamma}_k &= \mathbf{y}_{k-k_0} - \mathbf{C}\hat{\mathbf{x}}^F_{k|k-1} \\ &= \mathbf{C}\mathbf{x}_{k-k_0} + \mathbf{v}_{k-k_0} - \mathbf{C}\left(\mathbf{A}+\mathbf{BL}\right)\hat{\mathbf{x}}^F_{k-1|k-1} - \mathbf{CB}\mathbf{e}_{k-1}. \end{aligned} \tag{17}$$

Furthermore, let $\gamma_{e,k} = \left[\gamma_k^T, \mathbf{e}_{k-1}^T\right]^T$ and $\widetilde{\gamma}_{e,k} = \left[\widetilde{\gamma}_k^T, \mathbf{e}_{k-1}^T\right]^T$. Then, the PDFs before and after the attack will be as follows,

$f\left(\gamma_k, \mathbf{e}_{k-1}\right) = \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{\gamma_e}\right)$ and $\widetilde{f}\left(\widetilde{\gamma}_k, \mathbf{e}_{k-1}\right) = \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{\widetilde{\gamma}_e}\right)$,

where

$$\mathbf{\Sigma}_{\gamma_e} = \begin{bmatrix} \mathbf{\Sigma}_\gamma & \mathbf{0}_{m\times p} \\ \mathbf{0}_{p\times m} & \mathbf{\Sigma}_e \end{bmatrix} \text{ and } \mathbf{\Sigma}_{\widetilde{\gamma}_e} = \begin{bmatrix} \mathbf{\Sigma}_{\widetilde{\gamma}} & -\mathbf{CB}\mathbf{\Sigma}_e \\ -\mathbf{\Sigma}_e\mathbf{B}^T\mathbf{C}^T & \mathbf{\Sigma}_e \end{bmatrix}.$$

Also, $\gamma_k \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_\gamma\right)$ and $\widetilde{\gamma}_k \sim \mathcal{N}\left(\mathbf{0}, \mathbf{\Sigma}_{\widetilde{\gamma}}\right)$, where

$$\mathbf{\Sigma}_\gamma = \mathbf{C}\mathbf{P}^T\mathbf{C} + \mathbf{R}. \tag{18}$$

Here, $\mathcal{N}(\cdot)$ denotes normal distribution. $-\mathbf{CB}\mathbf{\Sigma}_e$ is the covariance matrix between $\widetilde{\gamma}_k$ and $\mathbf{e}_{k-1}$, which can be derived easy by multiplying $\mathbf{e}^T_{k-1}$ to both sides of (17), and then taking the expectation of both sides. Similarly, (18) can be derived by multiplying γ_k^T to both sides of (16), and then taking the expectation of both sides. In both the derivations, we need to utilise the appropriate information regarding uncorrelated variable pairs.

To derive the expression of $\mathbf{\Sigma}_{\widetilde{\gamma}}$, we need to evaluate the covariance of the attack signal, *i.e.*, $\mathbf{y}_{k-k_0}$, and the correlation between $\mathbf{x}_{k-1}$ and $\mathbf{y}_{k-k_0}$, see Lemma 1. Now, for the replay attack, the attack signal was generated by the same healthy system given in Subsection II-A k_0 time units ago. In order to simplify the derivation, we first transform the state-space representation of the original system given in (1)-(2) into a modified partially observed Gauss Markov process (GMP) as given in (19)-(20). In other words, (1)-(2) and (19)-(20) both represent the same system dynamics with different state-space definitions, and we can say that the attack signal $\mathbf{y}_{k-k_0}$ from (1)-(2) and the output $\mathbf{z}_k$ from (19)-(20)

both will have identical statistical properties. Therefore, instead of using $\mathbf{y}_{k-k_0}$, we derive the variance of $\mathbf{z}_k$, *i.e.*, $\mathbf{E}_{zz}(0)$, and the correlation between $\mathbf{x}_{k-1}$ and $\mathbf{z}_k$, *i.e.*, $\mathbf{E}_{xz}(-1)$, see Lemma 1. The relationships of the states and parameters between the two representations of the same system dynamics are derived as in (21)-(25), with the GMP being described as

$$\mathbf{x}_{a,k} = \mathbf{A}_a \mathbf{x}_{a,k-1} + \mathbf{w}_{a,k-1}, \tag{19}$$

$$\mathbf{z}_k = \mathbf{C}_a \mathbf{x}_{a,k}, \tag{20}$$

where $\mathbf{x}_{a,k} \in \mathbb{R}^{n_a}$ and $\mathbf{w}_{a,k} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_a)$ are the hidden state vector and iid noise vector, respectively, at the k-th time instant. $\mathbf{x}_{a,k}$, $\mathbf{w}_{a,k}$, $\mathbf{A}_a$, $\mathbf{C}_a$ and $\mathbf{Q}_a$ for the GMP will take the following forms,

$$\mathbf{x}_{a,k} = \begin{bmatrix} \mathbf{x}_{k-k_0} & \hat{\mathbf{x}}_{k-k_0|k-k_0-1} & \mathbf{v}_{k-k_0} \end{bmatrix}^T, \tag{21}$$

$$\mathbf{w}_{a,k} = \begin{bmatrix} \mathbf{B}\mathbf{e}_{k-k_0} + \mathbf{w}_{k-k_0} & \mathbf{B}\mathbf{e}_{k-k_0} & \mathbf{v}_{k-k_0+1} \end{bmatrix}^T, \tag{22}$$

$$\mathbf{A}_a = \begin{bmatrix} \mathbf{A}+\mathbf{BLKC} & \mathbf{BL}\left(\mathbf{I}_n - \mathbf{KC}\right) & \mathbf{BLK} \\ \left(\mathbf{A}+\mathbf{BL}\right)\mathbf{KC} & \left(\mathbf{A}+\mathbf{BL}\right)\left(\mathbf{I}_n - \mathbf{KC}\right) & \left(\mathbf{A}+\mathbf{BL}\right)\mathbf{K} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}, \tag{23}$$

$$\mathbf{C}_a = \begin{bmatrix} \mathbf{C} & \mathbf{0} & \mathbf{I}_n \end{bmatrix}, \tag{24}$$

$$\mathbf{Q}_a = \begin{bmatrix} \mathbf{B}\mathbf{\Sigma}_e\mathbf{B}^T + \mathbf{Q} & \mathbf{B}\mathbf{\Sigma}_e\mathbf{B}^T & \mathbf{0} \\ \mathbf{B}\mathbf{\Sigma}_e\mathbf{B}^T & \mathbf{B}\mathbf{\Sigma}_e\mathbf{B}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R} \end{bmatrix}. \tag{25}$$

$\mathbf{I}_n$ is the identity matrix of size n. The derivations follow directly from the comparison of the parameters of (1)-(2) and (19)-(20). Note that the state-space description (19)-(20) is simply a modelling exercise in representing $\mathbf{y}_{k-k_0}$, the attack signal, through a virtual linear state-space representation, allowing us to use the machinery of [27]. The attacker does not need to have access to this model, (19)-(20), and simply substitutes the true measurements with a previously recorded sequence.

Lemma 1. *The covariance matrix $\mathbf{\Sigma}_{\widetilde{\gamma}}$ of the innovation signal $\widetilde{\gamma}$ after the replay attack will take the following form for the attack model given in (19)-(20),*

$$\begin{aligned} \mathbf{\Sigma}_{\widetilde{\gamma}} = {} & \mathbf{E}_{zz}(0) - \mathbf{C}(\mathbf{A}+\mathbf{BL})\mathbf{E}_{xz}(-1) \\ & - \left[\mathbf{C}(\mathbf{A}+\mathbf{BL})\mathbf{E}_{xz}(-1)\right]^T + \mathbf{CB}\mathbf{\Sigma}_e\mathbf{B}^T\mathbf{C}^T \\ & + \mathbf{C}(\mathbf{A}+\mathbf{BL})\mathbf{\Sigma}_{x^F z}(\mathbf{A}+\mathbf{BL})^T\mathbf{C}^T \\ & + \mathbf{C}(\mathbf{A}+\mathbf{BL})\mathbf{\Sigma}_{x^F e}(\mathbf{A}+\mathbf{BL})^T\mathbf{C}^T, \end{aligned} \tag{26}$$

where

$$\mathbf{E}_{xz}(-1) = \sum_{i=0}^{\infty} \mathcal{A}^i \mathbf{K}\mathbf{C}_a\mathbf{A}_a^{i+1}\mathbf{E}_{x_a}(0)\,\mathbf{C}_a^T, \tag{27}$$

$\mathbf{E}_{zz}(0) = E\left[\mathbf{z}_k\mathbf{z}_k^T\right]$, $\mathbf{E}_{xz}(-1) = E\left[\mathbf{x}_{k-1}\mathbf{z}_k^T\right]$, *and* $\mathbf{E}_{x_a}(0) = E\left[\mathbf{x}_{a,k}\mathbf{x}_{a,k}^T\right]$. $\mathcal{A} = \left(\mathbf{I}_n - \mathbf{KC}\right)\left(\mathbf{A}+\mathbf{BL}\right)$. $\mathbf{\Sigma}_{x^F z}$ *and* $\mathbf{\Sigma}_{x^F e}$ *are the solutions to the following Lyapunov equations,*

$$\begin{aligned} & \mathcal{A}\mathbf{\Sigma}_{x^F z}\mathcal{A}^T - \mathbf{\Sigma}_{x^F z} + \mathbf{K}\mathbf{E}_{zz}(0)\mathbf{K}^T + \mathcal{A}\mathbf{E}_{xz}(-1)\mathbf{K}^T \\ & + \left(\mathcal{A}\mathbf{E}_{xz}(-1)\mathbf{K}^T\right)^T = 0, \text{ and} \end{aligned} \tag{28}$$

$$\mathcal{A}\mathbf{\Sigma}_{x^F e}\mathcal{A}^T - \mathbf{\Sigma}_{x^F e} + \left(\mathbf{I}_n - \mathbf{KC}\right)\mathbf{B}\mathbf{\Sigma}_e\mathbf{B}^T\left(\mathbf{I}_n - \mathbf{KC}\right)^T = 0. \tag{29}$$

Since $\mathcal{A}$ is assumed to be strictly stable, the Lyapunov equations (28) and (29) will have unique solutions [28].

Proof. The proof of Lemma 1 is provided in Appendix A. □

Based on the derivations that followed Proposition 1, we finally evaluate the CUSUM test statistics g_k at each time instant as follows,

$$g_k = \max\left(0, g_{k-1} + \log \frac{\mid \mathbf{\Sigma}_{\widetilde{\gamma}_e} \mid^{1/2} \exp\left(-0.5\bar{\gamma}_{e,k}^T \mathbf{\Sigma}_{\widetilde{\gamma}_e}^{-1} \bar{\gamma}_{e,k}\right)}{\mid \mathbf{\Sigma}_{\gamma_e} \mid^{1/2} \exp\left(-0.5\bar{\gamma}_{e,k}^T \mathbf{\Sigma}_{\gamma_e}^{-1} \bar{\gamma}_{e,k}\right)}\right). \tag{30}$$

Here, $\mid \cdot \mid$ denotes the determinant of a matrix. $\bar{\gamma}_{e,k}$ will be $\gamma_{e,k}$ before the attack, and after attack $\bar{\gamma}_{e,k} = \widetilde{\gamma}_{e,k}$.

B. Asymptotic Detection Performance

In this subsection, we study the asymptotic performance of the proposed replay attack detection scheme using supremum of ADD (SADD) and ARL as the two metrics. We analyse the quantities that affect the detector performance and alter a few of them by proper designing as discussed in Sub-section III-C and III-D, thus improving the performance.

The SADD is defined as

$$SADD \triangleq \sup_{1 \le \nu < \infty} E_\nu\left[T_{cs} - \nu | T_{cs} > \nu\right]. \tag{31}$$

Here $E_\nu[\cdot]$ denotes the expectation with respect to the distribution of the test data, *i.e.*, the innovation and the watermarking signals when the system is under attack. T_{cs} is the time instant of attack detection. SADD is asymptotically inversely proportional to the KLD, $D\left(\widetilde{f}, f\right)$, between the two distributions $\widetilde{f}(\cdot)$ and $f(\cdot)$, see Theorem 8.2.3 and Section 2 from [25] and [26], respectively, as follows,

$$SADD \to \frac{\log(ARL_h)}{D\left(\widetilde{f}, f\right)}, \text{ as } ARL_h \to \infty. \tag{32}$$

Here, $D\left(\widetilde{f}, f\right)$ takes the following form [27],

$$\begin{aligned} D\left(\widetilde{f}, f\right) = \frac{1}{2}\Big\{ & tr\left(\mathbf{\Sigma}_\gamma^{-1}\mathbf{\Sigma}_{\widetilde{\gamma}}\right) - m \\ & - \log \frac{\mid \mathbf{\Sigma}_{\widetilde{\gamma}} - \mathbf{CB}\mathbf{\Sigma}_e\mathbf{B}^T\mathbf{C}^T \mid}{\mid \mathbf{\Sigma}_\gamma \mid} \Big\}. \end{aligned} \tag{33}$$

The average run length is defined as, $ARL \triangleq E_\infty\left[T_{cs}\right]$, where $E_\infty[\cdot]$ denotes the expectation with respect to the distribution of the test data when no attack is present. From (32) and (33), we can comment that KLD and subsequently SADD are dependent on the watermarking signal variance $\mathbf{\Sigma}_e$. We have used the KLD expression in (33) to find the optimal $\mathbf{\Sigma}_e$ that maximizes the KLD for a given upper bound on the increase in the control cost in the following Subsection III-C.

Remark 2. *As given in Corollary 2.1 in [27], the KLD for the suboptimal CUSUM case is lower compared to that of the optimal CUSUM case. Such reduction in KLD increases the detection delay for the suboptimal CUSUM case.*

C. Optimal watermarking signal variance

The addition of watermarking increases the KLD, but at the same time, it also increases the control cost. The increase in LQG

control cost, ΔLQG, due to the addition of watermarking is given in [27] as follows,

$$\Delta LQG = tr\left[\left(\mathbf{B}^T\mathbf{\Sigma}_L\mathbf{B}+\mathbf{U}\right)\mathbf{\Sigma}_e\right], \tag{34}$$

where $\mathbf{\Sigma}_L$ is the solution to the Lyapunov equation

$$\left(\mathbf{A}+\mathbf{BL}\right)^T\mathbf{\Sigma}_L\left(\mathbf{A}+\mathbf{BL}\right)-\mathbf{\Sigma}_L+\mathbf{L}^T\mathbf{UL}+\mathbf{W}=0. \tag{35}$$

Therefore, we want to find the optimal $\mathbf{\Sigma}_e$ that will maximize the KLD for a given fixed threshold J on the ΔLQG. According to the Theorem 5 from [27], the optimal $\mathbf{\Sigma}_e$ will have only one non-zero eigenvalue. Therefore, we search for the optimum $\mathbf{\Sigma}_e$ within the class of rank one positive semi-definite matrices with the following structure,

$$\mathbf{\Sigma}_e = \mathbf{v}_\lambda\mathbf{v}_\lambda^T, \text{ where } \mathbf{v}_\lambda = \sqrt{\lambda}_e\mathbf{v}_e. \tag{36}$$

Here, λ_e is the non-zero eigenvalue and $\mathbf{v}_e$ is the corresponding eigenvector. Now, we define the optimization problem as,

$$\begin{aligned}&\max_{\mathbf{v}_\lambda} D\left(\widetilde{f},f\right)\\ &\text{s.t. } \Delta LQG \le J.\end{aligned} \tag{37}$$

We have solved the optimization problem using the interior point method [29]. It can also be solved by other non-convex optimizers, such as sequential quadratic programming (SQP) [30], etc. Since the cost function is non-concave, the solution may only be a local optimum.

D. Systems with High Relative Degree

The system given in (1)-(2) is said to have relative degree d_r provided [31]

$$\begin{aligned}&\mathbf{CA}^i\mathbf{B}=\mathbf{0}, \text{ for } i<d_r-1, \text{ and}\\ &\mathbf{CA}^i\mathbf{B}\neq\mathbf{0}, \text{ for } i=d_r-1.\end{aligned} \tag{38}$$

If the system under consideration has a relative degree $d_r = k_e$, where $k_e \ge 2$, then $\mathbf{CB}$ becomes $\mathbf{0}$. Therefore, for such systems, the $\mathbf{CB\Sigma}_e\mathbf{B}^T\mathbf{C}^T$ term in the KLD expression (33) will vanish, which will reduce the overall KLD. In such a situation, the joint distribution of the innovation signal γ_k or $\widetilde{\gamma}_k$ and the delayed version of the watermarking signal, *i.e.*, $\mathbf{e}_{k-k_e}$, can improve the KLD. Increase in KLD means faster attack detection. The proposed CUSUM test and the corresponding KLD using the joint distribution of the innovation signal and the delayed watermarking signal $\mathbf{e}_{k-k_e}$ are provided in the following theorem.

Theorem 1. *If the system has a relative degree of $d_r = k_e$, and the joint distribution of the innovation signal and the watermarking signal $\mathbf{e}_{k-k_e}$ is considered for the CUSUM test, then the test statistics, $g_{k_e,k}$, and the KLD, $D_d\left(\widetilde{f}_{k_e},f_{k_e}\right)$, between the normal system and the system under attack will be as follows*

$$g_{k_e,k} = \max\left(0, g_{k_e,k-1}+\log\frac{\widetilde{f}_{k_e}\left(\bar{\gamma}_k,\mathbf{e}_{k-k_e}\right)}{f_{k_e}\left(\bar{\gamma}_k,\mathbf{e}_{k-k_e}\right)}\right) \tag{39}$$

$$\begin{aligned}D_d\left(\widetilde{f}_{k_e},f_{k_e}\right) &= \frac{1}{2}\Big\{tr\left(\mathbf{\Sigma}_\gamma^{-1}\mathbf{\Sigma}_{\widetilde{\gamma}}\right)-m\\ &-\log\frac{\mid\mathbf{\Sigma}_{\widetilde{\gamma}}-\mathbf{CA}^{k_e-1}\mathbf{B\Sigma}_e\mathbf{B}^T(\mathbf{A}^{k_e-1})^T\mathbf{C}^T\mid}{\mid\mathbf{\Sigma}_\gamma\mid}\Big\},\end{aligned} \tag{40}$$

where $f_{k_e}(\cdot)$ and $\widetilde{f}_{k_e}(\cdot)$ denote the joint probability density functions (PDF) of innovation signal and $\mathbf{e}_{k-k_e}$, before and after the attack, respectively, see (41) and (42). The expressions for $\mathbf{\Sigma}_\gamma$ and $\mathbf{\Sigma}_{\widetilde{\gamma}}$ are the same as in (18) and (26), respectively. Let $\gamma_{e,k}^{k_e} = \left[\gamma_k^T,\mathbf{e}_{k-k_e}^T\right]^T$ and $\widetilde{\gamma}_{e,k}^{k_e} = \left[\widetilde{\gamma}_k^T,\mathbf{e}_{k-k_e}^T\right]^T$. Then, the PDFs before and after the attack will be as follows,

$$f_{k_e}\left(\gamma_k,\mathbf{e}_{k-k_e}\right) = \mathcal{N}\left(\mathbf{0},\mathbf{\Sigma}_{\gamma_e^{k_e}}\right) \text{ and} \tag{41}$$

$$\widetilde{f}_{k_e}\left(\widetilde{\gamma}_k,\mathbf{e}_{k-k_e}\right) = \mathcal{N}\left(\mathbf{0},\mathbf{\Sigma}_{\widetilde{\gamma}_e^{k_e}}\right), \text{where} \tag{42}$$

$$\mathbf{\Sigma}_{\gamma_e^{k_e}} = \mathbf{\Sigma}_{\gamma_e} \text{ and} \tag{43}$$

$$\mathbf{\Sigma}_{\widetilde{\gamma}_e^{k_e}} = \begin{bmatrix}\mathbf{\Sigma}_{\widetilde{\gamma}} & -\mathbf{CA}^{k_e-1}\mathbf{B\Sigma}_e\\ -\mathbf{\Sigma}_e\mathbf{B}^T\left[\mathbf{A}^{k_e-1}\right]^T\mathbf{C}^T & \mathbf{\Sigma}_e\end{bmatrix}. \tag{44}$$

Proof. To prove Theorem 1, we only need to show that $\mathrm{E}\left[\gamma_k\mathbf{e}_{k-k_e}^T\right] = \mathbf{0}$ and $\mathrm{E}\left[\widetilde{\gamma}_k\mathbf{e}_{k-k_e}^T\right] = -\mathbf{CA}^{k_e-1}\mathbf{B\Sigma}_e$, which are derived in Appendix B. □

Remark 3. *The expression of $\mathbf{\Sigma}_{\widetilde{\gamma}}$ (26) can be simplified using the information $\mathbf{CB}=0$ for a system with $d_r \ge 2$.*

E. CUSUM vs. NP Detector

In this paper, one of our research objectives is to compare the proposed sequential detection based method (CUSUM) with a member of non-sequential based methods. The optimal NP based χ^2 detector used for the comparison is a non-sequential detector. The compared method was first introduced in 2009 [32], and then improved upon several years [3]. The NP based χ^2 detector is considered to be one of the state-of-the-art methods for replay attack detections in dynamical systems. On the other hand, the CUSUM test is the quickest sequential detector in the sense that it minimizes the worst-case ADD, i.e., SADD for a fixed lower limit on ARL [25]. That means the CUSUM test will perform better than the non-sequential based methods such as NP based χ^2 detectors in terms of the average detection delay. In this paper, we have ignored the dependency of $\widetilde{\gamma}_k$ on its previous values, which makes the corresponding CUSUM test a non-optimal CUSUM. However, we have shown by simulation in Sub-section IV-D that the proposed non-optimal CUSUM based replay attack detector is performing better than the NP based χ^2 detector [3].

The watermarking signal is taken to be iid, and the $\mathbf{\Sigma}_e$ is optimized for both cases. In [3], the optimal NP detector rejects the H_0 hypothesis in favour of H_1 if

$$\begin{aligned}&g_{NP,k}\left(\gamma_k,\mathbf{e}_{k-1},\cdots\right) = \gamma_k^T\mathbf{\Sigma}_\gamma^{-1}\gamma_k\\ &-\left(\gamma_k-\mu_{NP,k}\right)^T\left(\mathbf{\Sigma}_\gamma+\mathbf{\Sigma}_f\right)^{-1}\left(\gamma_k-\mu_{NP,k}\right)\ge\eta,\end{aligned} \tag{45}$$

$$\text{where } \mu_{NP,k} = -\mathbf{C}\sum_{i=-\infty}^{k}\mathcal{A}^{k-i}\mathbf{B}\mathbf{e}_i, \tag{46}$$

$$\mathbf{\Sigma}_f = \mathbf{C}\mathcal{L}_f\mathbf{C}^T, \text{ and} \tag{47}$$

$$\mathcal{L}_f = \mathcal{A}\mathcal{L}_f\mathcal{A}^T+\mathbf{B\Sigma}_e\mathbf{B}^T. \tag{48}$$

The ADD is estimated as,

$$ADD_{NP} = E\left[\inf\left\{k : g_{NP,k}(\cdot)\ge\eta\right\}\right], \tag{49}$$

where η is the user selected threshold. To compare the methods under study on the same ground, we have derived the thresholds for the test statistics for both the tests by keeping a fixed lower limit on ARL, *i.e.*, ARL_{th}. The proposed CUSUM test inherently uses ARL_{th} information (15) to derive the threshold for the test statistics. However, for the NP-based method, the threshold η is derived from the Monte-Carlo simulation using $ARL \ge ARL_{th}$

as a constraint, and selecting the lowest feasible threshold, as to maximise the probability of detection.

IV. Numerical Results

In this section, we illustrate the replay attack detection methodology proposed in this paper using three different system models. The three different systems are System-A: A second-order open-loop unstable multiple inputs and single output (MISO) system, System-B: A fourth-order open-loop stable multiple inputs and multiple outputs (MIMO) system, and System-C: A second-order open-loop unstable MISO system with relative degree two. System-B is a linearized minimum phase quadruple tank system taken from [33]. Only the level sensor gains are increased to make the magnitude of the product $\mathbf{CB}$ numerically significant. The system parameters are provided as follows.
For the System-A and System-B, $ARL_h = 1000$. For the System-C, $ARL_h = 100$.

System-A parameters:

$$\mathbf{A} = \begin{bmatrix} 0.75 & 0.2 \\ 0.2 & 1.0 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0.9 & 0.5 \\ 0.1 & 1.2 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 1.0 & -1.0 \end{bmatrix}$$
$$\mathbf{Q} = diag\begin{bmatrix} 1 & 1 \end{bmatrix} \quad \mathbf{R} = 1 \quad \mathbf{W} = diag\begin{bmatrix} 1 & 2 \end{bmatrix}$$
$$\mathbf{U} = diag\begin{bmatrix} 0.4 & 0.7 \end{bmatrix}$$

System-B parameters:

$$\mathbf{A} = \begin{bmatrix} 0.9683 & 0 & 0.0819 & 0 \\ 0 & 0.9780 & 0 & 0.06377 \\ 0 & 0 & 0.9167 & 0 \\ 0 & 0 & 0 & 0.9355 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 0.1638 & 0.004 \\ 0.002 & 0.1242 \\ 0 & 0.0917 \\ 0.0604 & 0 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \end{bmatrix}$$
$$\mathbf{Q} = diag\begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix} \quad \mathbf{R} = diag\begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$$
$$\mathbf{W} = diag\begin{bmatrix} 5 & 5 & 1 & 1 \end{bmatrix} \quad \mathbf{U} = diag\begin{bmatrix} 2 & 2 \end{bmatrix}$$

System-C parameters:

$$\mathbf{B} = \begin{bmatrix} 0.9 & 0.5 \\ 1.3 & 0.72 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 1.3 & -0.9 \end{bmatrix}$$

The rest of the parameters are the same as System-A.
To evaluate the SADD for each ΔLQG value, first, we have incremented the attack start point ν from 1 to 1000 with a step size of 1. For each attack start point ν, we have estimated the ADD over 1000 Monte-Carlo trials. Finally, we have evaluated the SADD as the highest ADD (worse case ADD) over the range of ν.

A. Replay attack detection

Figure 3 shows the trade-off between the SADD and the increase in ΔLQG when the System-B is under a replay attack. We plot the derived SADD using the theory developed in this paper, and the estimated SADD using the simulated data, where $\mathbf{\Sigma}_e$ is assumed to be diagonal and all the watermarking signals have equal power. Therefore, we can detect an attack early at the expense of an increased control cost. Figure 3 also illustrates that it is hard to detect a replay attack with low watermarking signal power. This is implicit in SADD, but there is a sharp increase observed before $\Delta LQG \approx 0.8$, which corresponds to $\Sigma_e = \text{diag}\begin{bmatrix} 0.29 & 0.29 \end{bmatrix}$. However, the sharp corner point at $\Delta LQG \approx 0.8$ is the effect of limited data points on the SADD vs ΔLQG curve.

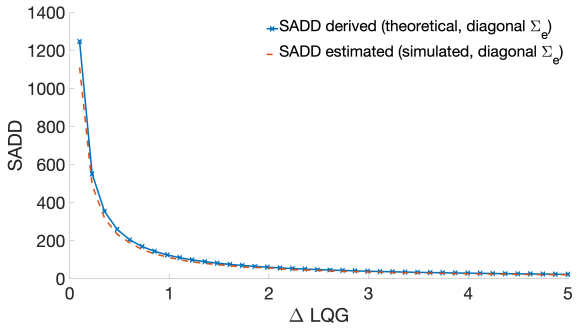

Figure 3: SADD vs. ΔLQG plot for System-B under replay attack.

B. Optimum and non-optimum $\mathbf{\Sigma}_e$

Figure 4 shows the SADD vs ΔLQG plots for System-A using the optimized $\mathbf{\Sigma}_e$ and a diagonal $\mathbf{\Sigma}_e$ with equal signal power when the system is under a replay attack. The optimum LQG value for the system before the attack and without the added watermarking is 38.03. It is evident that optimizing $\mathbf{\Sigma}_e$ improves the SADD for a fixed upper threshold on ΔLQG. We can also say that the same SADD can be achieved for a much reduced ΔLQG.

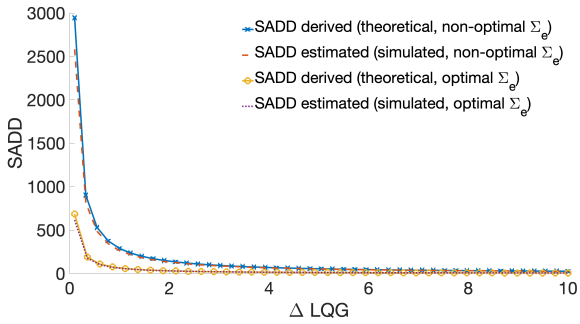

Figure 4: SADD vs. ΔLQG plot for System-A under replay attack with optimum and non-optimum $\mathbf{\Sigma}_e$.

C. System with higher relative degree

Figure 5 shows the benefit of using the delayed version of watermarking signal, *i.e.*, $\mathbf{e}_{k-k_e}$ for a system with relative degree $d_r = k_e$ as discussed in Theorem 1. System-C with relative degree $d_r = 2$ is used to generate the plots of Fig. 5. We can see reductions in ΔLQG to achieve the same $SADD$ between any two points on the ΔLQG axis.

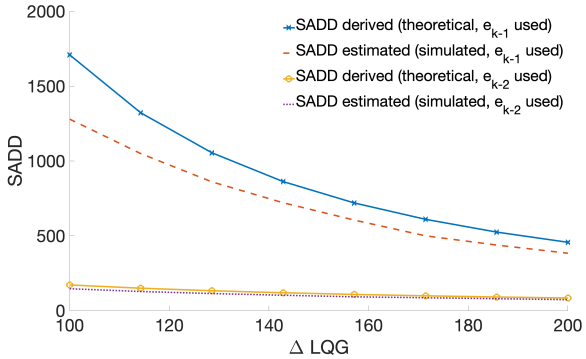

Figure 5: SADD vs. ΔLQG plots for System-C with relative degree 2.

D. Comparison with optimal NP-based detector

The left axis of Fig. 6 shows the tradeoff between the ADD and the increase in ΔLQG for System-A under the proposed CUSUM

test and the NP-based method reported in [3] (see Remark-II) for the detection of replay attacks. We plot the derived SADD using the theory developed in this paper, the estimated SADD applying the sub-optimal CUSUM test on the simulated data, and the estimated ADD applying the test reported in [3] on the simulated data using optimum $\mathbf{\Sigma}_e$. The higher ΔLQG portion of the plot is zoomed. The right axis of Fig. 6 shows the corresponding simulated ARLs for both the tests. It is clear from the figure that we can achieve lower detection delay for the same LQG loss with the method proposed in this paper compared to the one reported in [3]. ARL increases with the ΔLQG, *i.e.*, watermarking signal power, for the proposed method, whereas it does not change much for the method reported in [3].

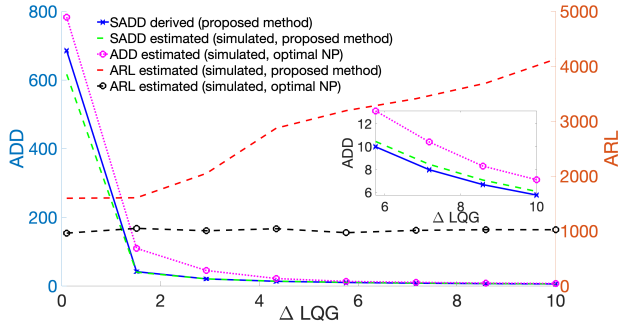

Figure 6: ADD and ARL vs. ΔLQG plot for System-A under proposed CUSUM test and NP test.

V. Conclusion

We have addressed the problem of resilient replay attack detection using the CUSUM test. The detection delay and the corresponding increase in the control cost are studied. The KLD expression between the distributions before the attack and after the replay attack is derived. The KLD reduces for the systems with relative-degree higher than one. We have proposed a technique of using a delayed version of the watermarking to improve the KLD for such systems. The simulation results shown are in close agreement with the theory presented in the paper. We have also discussed a way to optimize the watermarking signal variance to maximize the KLD under the replay attack for a fixed increase in the control cost.

Appendix A
Derivation of $\mathbf{\Sigma}_{\widetilde{\gamma}}$

Since the measurement $\mathbf{y}_{k-k_0}$ from (1)-(2) is stationary, $\mathbf{z}_k$ from (19)-(20) will also be stationary, since their statistical properties are identical. Therefore, the initial state covariance $\mathbf{E}_{x_a}(0)$ will be a constant, given by the solution to the following Lyapunov equation.

$$\mathbf{E}_{x_a}(0) = \mathbf{A}_a \mathbf{E}_{x_a}(0) \mathbf{A}_a^T + \mathbf{Q}_a. \tag{50}$$

The expression of $\mathbf{\Sigma}_{\widetilde{\gamma}} = E\left[\widetilde{\gamma}_k \widetilde{\gamma}_k^T\right]$ is derived as follows. Using (17), and applying the knowledge that $\mathbf{e}_{k-1}$ is uncorrelated with $\mathbf{z}_k$ and $\hat{\mathbf{x}}^F_{k-1|k-1}$, we get the following expression of $\mathbf{\Sigma}_{\widetilde{\gamma}}$,

$$\begin{aligned}\mathbf{\Sigma}_{\widetilde{\gamma}} &= E\left[\mathbf{z}_k \mathbf{z}_k^T\right] - \mathbf{C}\left(\mathbf{A}+\mathbf{BL}\right) E\left[\hat{\mathbf{x}}^F_{k-1|k-1} \mathbf{z}_k^T\right] \\ &- \left(\mathbf{C}\left(\mathbf{A}+\mathbf{BL}\right) E\left[\hat{\mathbf{x}}^F_{k-1|k-1} \mathbf{z}_k^T\right]\right)^T + \mathbf{CB}\mathbf{\Sigma}_e \mathbf{B}^T \mathbf{C}^T \\ &+ \mathbf{C}\left(\mathbf{A}+\mathbf{BL}\right) E\left[\hat{\mathbf{x}}^F_{k-1|k-1}\left(\hat{\mathbf{x}}^F_{k-1|k-1}\right)^T\right] \left(\mathbf{A}+\mathbf{BL}\right)^T \mathbf{C}^T.\end{aligned} \tag{51}$$

$E\left[\hat{\mathbf{x}}^F_{k-1|k-1} \mathbf{z}_k^T\right]$ is calculated as follows. First, using (12), we evaluate

$$\hat{\mathbf{x}}^F_{k-1|k-1} = \mathbf{K}\mathbf{z}_{k-1} + \mathcal{A}\hat{\mathbf{x}}^F_{k-2|k-2} + \left(\mathbf{I}_n - \mathbf{KC}\right)\mathbf{B}\mathbf{e}_{k-2}, \tag{52}$$

where $\mathcal{A} = \left(\mathbf{I}_n - \mathbf{KC}\right)\left(\mathbf{A}+\mathbf{BL}\right)$. We define

$$\begin{aligned}\mathbf{E}_{xz}(-k_0) &\triangleq E\left[\hat{\mathbf{x}}^F_{k-k_0|k-k_0} \mathbf{z}_k^T\right] = E\left[\left(\mathbf{K}\mathbf{z}_{k-k_0} + \right.\right. \\ &\left.\left. \mathcal{A}\hat{\mathbf{x}}^F_{k-k_0-1|k-k_0-1} + \left(\mathbf{I}_n - \mathbf{KC}\right)\mathbf{B}\mathbf{e}_{k-k_0-1}\right)\mathbf{z}_k^T\right], \text{ [using (52)]} \\ &= \mathbf{K}\mathbf{E}_{zz}(-k_0) + \mathcal{A}\mathbf{E}_{xz}(-k_0-1),\end{aligned} \tag{53}$$

where $\mathbf{e}_{k-k_0-1}$ and $\mathbf{z}_k$ are uncorrelated, and $\mathbf{E}_{zz}(-k_0) = E\left[\mathbf{z}_{k-k_0}\mathbf{z}_k^T\right]$. $\mathbf{E}_{x_a}(-k_0) = \mathbf{E}_{x_a}(k_0) \triangleq E\left[\mathbf{x}_{a,k}\mathbf{x}_{a,k-k_0}^T\right]$ is evaluated as

$$\begin{aligned}&\mathbf{E}_{x_a}(-1) = \mathbf{A}_a \mathbf{E}_{x_a}(0), \text{ [}\mathbf{w}_{a,k} \text{ and } \mathbf{x}_{a,k} \text{ uncorrelated].} \\ &\text{Similarly, } \mathbf{E}_{x_a}(-2) = \mathbf{A}_a \mathbf{E}_{x_a}(-1) = \mathbf{A}_a^2 \mathbf{E}_{x_a}(0), \text{ and} \\ &\mathbf{E}_{x_a}(-k_0) = \mathbf{A}_a^{k_0} \mathbf{E}_{x_a}(0).\end{aligned} \tag{54}$$

The system matrix $\mathbf{A}_a$ is assumed to be stable because the attacker will always try to generate fake observations which are bounded and will mimic the true observations to remain stealthy. For a stable $\mathbf{A}_a$, $\mathbf{A}_a^{k_0} \to 0$, as $k_0 \to \infty$. Therefore,

$$\mathbf{E}_{x_a}(-k_0) \to 0, \text{ as } k_0 \to \infty. \tag{55}$$

The expression for $\mathbf{E}_{zz}(-k_0) = \mathbf{E}_{zz}(k_0) \triangleq E\left[\mathbf{z}_{k-k_0}\mathbf{z}_k^T\right]$ is derived using (20) and (55) as

$$\mathbf{E}_{zz}(-k_0) = \mathbf{C}_a \mathbf{E}_{x_a}(-k_0)\mathbf{C}_a^T = \mathbf{C}_a \mathbf{A}_a^{k_0} \mathbf{E}_{x_a}(0)\mathbf{C}_a^T \tag{56}$$

Using (53) and (56), we can write the expression of $\mathbf{E}_{xz}(-1)$ as

$$\begin{aligned}&\mathbf{E}_{xz}(-1) = \mathbf{K}\mathbf{E}_{zz}(-1) + \mathcal{A}\mathbf{E}_{xz}(-2) \\ &= \mathbf{K}\mathbf{C}_a\mathbf{A}_a\mathbf{E}_{x_a}(0)\mathbf{C}_a^T + \mathcal{A}\mathbf{K}\mathbf{C}_a\mathbf{A}_a^2\mathbf{E}_{x_a}(0)\mathbf{C}_a^T + \mathcal{A}^2\mathbf{E}_{xz}(-3).\end{aligned} \tag{57}$$

Repeating the same technique, $\mathbf{E}_{xz}(-1)$ will take the following form,

$$\mathbf{E}_{xz}(-1) = \sum_{i=0}^{\infty} \mathcal{A}^i \mathbf{K}\mathbf{C}_a \mathbf{A}_a^{i+1} \mathbf{E}_{x_a}(0)\mathbf{C}_a^T. \tag{58}$$

$\mathbf{E}_{xz}(-1)$ can be evaluated numerically by taking a large number of terms for the summation (58), until the rest of the terms become negligible. Using (52), $\mathbf{E}_{x^F x^F}(0) \triangleq E\left[\hat{\mathbf{x}}^F_{k-1|k-1}\left(\hat{\mathbf{x}}^F_{k-1|k-1}\right)^T\right]$ is evaluated as follows.

$$\begin{aligned}\mathbf{E}_{x^F x^F}(0) &= \mathbf{K}E\left[\mathbf{z}_{k-1}\mathbf{z}_{k-1}^T\right]\mathbf{K}^T + \mathcal{A}E\left[\hat{\mathbf{x}}^F_{k-2|k-2}\mathbf{z}_{k-1}^T\right]\mathbf{K}^T \\ &+ \left(\mathcal{A}E\left[\hat{\mathbf{x}}^F_{k-2|k-2}\mathbf{z}_{k-1}^T\right]\mathbf{K}^T\right)^T + \mathcal{A}E\left[\hat{\mathbf{x}}^F_{k-2|k-2}\left(\hat{\mathbf{x}}^F_{k-2|k-2}\right)^T\right]\mathcal{A}^T \\ &+ \left(\mathbf{I}_n - \mathbf{KC}\right)\mathbf{B}E\left[\mathbf{e}_{k-2}\mathbf{e}_{k-2}^T\right]\mathbf{B}^T\left(\mathbf{I}_n - \mathbf{KC}\right)^T.\end{aligned}$$

Therefore, $\mathbf{E}_{x^F x^F}(0)$ is the solution to the following Lyapunov equation,

$$\begin{aligned}&\mathcal{A}\mathbf{E}_{x^F x^F}(0)\mathcal{A}^T - \mathbf{E}_{x^F x^F}(0) + \mathbf{K}\mathbf{E}_{zz}(0)\mathbf{K}^T \\ &+ \mathcal{A}\mathbf{E}_{xz}(-1)\mathbf{K}^T + \left(\mathcal{A}\mathbf{E}_{xz}(-1)\mathbf{K}^T\right)^T + \\ &\left(\mathbf{I}_n - \mathbf{KC}\right)\mathbf{B}\mathbf{\Sigma}_e\mathbf{B}^T\left(\mathbf{I}_n - \mathbf{KC}\right)^T = 0, \text{ [(56) used].}\end{aligned} \tag{59}$$

$\mathbf{E}_{x^F x^F}(0)$ is divided into two parts, $\mathbf{\Sigma}_{x^F z}$ and $\mathbf{\Sigma}_{x^F e}$ which are independent of the watermarking signal and the fake observations,

respectively. $\mathbf{\Sigma}_{x^F z}$ and $\mathbf{\Sigma}_{x^F e}$ are the solution to the following Lyapunov equations,

$$\begin{aligned}
&\mathcal{A}\mathbf{\Sigma}_{x^F z}\mathcal{A}^T - \mathbf{\Sigma}_{x^F z} + \mathbf{K}\mathbf{E}_{zz}(0)\mathbf{K}^T + \mathcal{A}\mathbf{E}_{xz}(-1)\mathbf{K}^T \\
&+ \left(\mathcal{A}\mathbf{E}_{xz}(-1)\mathbf{K}^T\right)^T = 0, \\
&\mathcal{A}\mathbf{\Sigma}_{x^F e}\mathcal{A}^T - \mathbf{\Sigma}_{x^F e} + (\mathbf{I}_n - \mathbf{K}\mathbf{C})\,\mathbf{B}\mathbf{\Sigma}_e\mathbf{B}^T\,(\mathbf{I}_n - \mathbf{K}\mathbf{C})^T = 0, \\
&\text{and } \mathbf{E}_{x^F x^F}(0) = \mathbf{\Sigma}_{x^F z} + \mathbf{\Sigma}_{x^F e}. \qquad (60)
\end{aligned}$$

Using (56) and (60), we can rewrite the expression for $\mathbf{\Sigma}_{\widetilde{\gamma}}$ as given in (26).

Appendix B
Proof of Theorem 1

Assumption: System has relative degree $d_r = k_e$. Since γ_k before the attack is iid, $E\left[\gamma_k \mathbf{e}_{k-k_e}^T\right] = \mathbf{0}$. Applying (38) in (17), and replacing $\mathbf{y}_{k-k_0}$ by $\mathbf{z}_k$, since they both have same statistical properties, we get,

$$\widetilde{\gamma}_k = \mathbf{z}_k - \mathbf{C}\mathbf{A}\hat{\mathbf{x}}_{k-1|k-1}^F. \qquad (61)$$

$$\text{Therefore, } E\left[\widetilde{\gamma}_k \mathbf{e}_{k-k_e}^T\right] = -\mathbf{C}\mathbf{A}E\left[\hat{\mathbf{x}}_{k-1|k-1}^F \mathbf{e}_{k-k_e}^T\right], \qquad (62)$$

$$\text{and } \hat{\mathbf{x}}_{k-1|k-1}^F = \mathbf{K}\mathbf{z}_{k-1} + \mathcal{A}\hat{\mathbf{x}}_{k-2|k-2}^F + \mathbf{B}\mathbf{e}_{k-2}, \qquad (63)$$

where $\mathbf{e}_{k-k_e}$ is uncorrelated to $\mathbf{z}_k$. Using (63) recursively, we derive

$$\hat{\mathbf{x}}_{k-1|k-1}^F = \sum_{i=2}^{k_e}\left(\mathcal{A}^{i-2}\mathbf{K}\bar{\mathbf{y}}_{k-i+1} + \mathcal{A}^{i-2}\mathbf{B}\mathbf{e}_{k-i}\right) + \qquad (64)$$

$$\mathcal{A}^{k_e-1}\hat{\mathbf{x}}_{k-k_e|k-k_e}^F \left[\bar{\mathbf{y}}_k = \mathbf{y}_k \text{ if } k < \nu, \text{ and } \bar{\mathbf{y}}_k = \mathbf{z}_k \text{ otherwise.}\right]$$

Applying (64) in (62) and using (38), we get

$$E\left[\widetilde{\gamma}_k \mathbf{e}_{k-k_e}^T\right] = -\mathbf{C}\mathbf{A}\mathcal{A}^{k_e-2}\mathbf{B}\mathbf{\Sigma}_e \qquad (65)$$

where $\mathbf{e}_{k-k_e}$ is uncorrelated to $\hat{\mathbf{x}}_{k-k_e|k-k_e}^F$ and $\mathbf{z}_{k-i+1}$. Applying multinomial theorem on $\mathcal{A}^{k_e-1} = (\mathbf{A} - \mathbf{K}\mathbf{C}\mathbf{A} + \mathbf{B}\mathbf{L})^{k_e-1}$ and using (38), we get

$$E\left[\widetilde{\gamma}_k \mathbf{e}_{k-k_e}^T\right] = -\mathbf{C}\mathbf{A}^{k_e-1}\mathbf{B}\mathbf{\Sigma}_e. \qquad (66)$$

References

[1] R. Alguliyev, Y. Imamverdiyev, and L. Sukhostat, "Cyber-physical systems and their security issues," *Comput. Ind.*, vol. 100, no. July 2017, pp. 212–223, 2018.

[2] R. Langner, "Stuxnet: Dissecting a cyberwarfare weapon," *IEEE Secur. Priv.*, vol. 9, no. 3, pp. 49–51, 2011.

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

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

[5] D. Ding, Q. L. Han, Y. Xiang, X. Ge, and X. M. Zhang, "A survey on security control and attack detection for industrial cyber-physical systems," *Neurocomputing*, vol. 275, pp. 1674–1683, 2018.

[6] Y. Zhao and C. Smidts, "A control-theoretic approach to detecting and distinguishing replay attacks from other anomalies in nuclear power plants," *Prog. Nucl. Energy*, vol. 123, no. March, p. 103315, 2020.

[7] M. Hosseinzadeh, B. Sinopoli, and E. Garone, "Feasibility and Detection of Replay Attack in Networked Constrained Cyber-Physical Systems," *2019 57th Annu. Allert. Conf. Commun. Control. Comput. Allert. 2019*, pp. 712–717, 2019.

[8] L. Zhai and K. G. Vamvoudakis, "A data-based private learning framework for enhanced security against replay attacks in cyber-physical systems," *Int. J. Robust Nonlinear Control*, no. January, pp. 1–17, 2020.

[9] H. Liu, Y. Mo, J. Yan, L. Xie, and K. H. Johansson, "An online approach to physical watermark design," *IEEE Transactions on Automatic Control*, vol. 65, no. 9, pp. 3895–3902, 2020.

[10] 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.

[11] ——, "A switching multiplicative watermarking scheme for detection of stealthy cyber-attacks," *IEEE Transactions on Automatic Control*, vol. 66, no. 6, pp. 2558–2573, 2020.

[12] H. S. Sánchez, D. Rotondo, T. Escobet, V. Puig, J. Saludes, and J. Quevedo, "Detection of replay attacks in cyber-physical systems using a frequency-based signature," *J. Franklin Inst.*, vol. 356, no. 5, pp. 2798–2824, 2019.

[13] D. Ye, T. Y. Zhang, and G. Guo, "Stochastic coding detection scheme in cyber-physical systems against replay attack," *Inf. Sci. (Ny).*, vol. 481, no. 61773097, pp. 432–444, 2019.

[14] 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, 2020.

[15] D. Shi, Z. Guo, K. H. Johansson, and L. Shi, "Causality countermeasures for anomaly detection in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 63, no. 2, pp. 386–401, 2017.

[16] M. Porter, P. Hespanhol, A. Aswani, M. Johnson-Roberson, and R. Vasudevan, "Detecting Generalized Replay Attacks via Time-Varying Dynamic Watermarking," *IEEE Trans. Automat. Contr.*, vol. 66, no. 8, pp. 1–1, 2020.

[17] S. Rath, D. Pal, P. S. Sharma, and B. K. Panigrahi, "A Cyber-Secure Distributed Control Architecture for Autonomous AC Microgrid," *IEEE Syst. J.*, pp. 1–12, 2020.

[18] A. Hoehn and P. Zhang, "Detection of replay attacks in cyber-physical systems," *Proc. Am. Control Conf.*, vol. 2016-July, pp. 290–295, 2016.

[19] L. Liu, L. Ma, Y. Wang, J. Zhang, and Y. Bo, "Distributed set-membership filtering for time-varying systems under constrained measurements and replay attacks," *J. Franklin Inst.*, vol. 357, no. 8, pp. 4983–5003, 2020.

[20] J. Huang, L. Zhao, and Q. G. Wang, "Adaptive control of a class of strict feedback nonlinear systems under replay attacks," *ISA Trans.*, vol. 107, pp. 134–142, 2020.

[21] A. J. Gallo, M. S. Turan, F. Boem, G. Ferrari-Trecate, and T. Parisini, "Distributed watermarking for secure control of microgrids under replay attacks," *IFAC-PapersOnLine*, vol. 51, no. 23, pp. 182–187, 2018.

[22] G. Franze, F. Tedesco, and W. Lucia, "Resilient Control for Cyber-Physical Systems Subject to Replay Attacks," *IEEE Control Syst. Lett.*, vol. 3, no. 4, pp. 984–989, 2019.

[23] B. Chen, D. W. Ho, G. Hu, and L. Yu, "Secure Fusion Estimation for Bandwidth Constrained Cyber-Physical Systems under Replay Attacks," *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1862–1876, 2018.

[24] J. Giraldo and A. A. Cardenas, "A new metric to compare anomaly detection algorithms in cyber-physical systems," in *Proc. 6th Annu. Symp. Hot Top. Sci. Secur.*, 2019, pp. 1–2.

[25] A. Tartakovsky, I. Nikiforov, and M. Basseville, *Sequential analysis: Hypothesis testing and changepoint detection*, 2014.

[26] V. Girardin, V. Konev, and S. Pergamenchtchikov, "Kullback-Leibler Approach to CUSUM Quickest Detection Rule for Markovian Time Series," *Seq. Anal.*, vol. 37, no. 3, pp. 322–341, 2018.

[27] A. Naha, A. Teixeira, A. Ahlén, and S. Dey, "Quickest detection of deception attacks in networked control systems with physical watermarking," *arXiv preprint arXiv:2101.01466*, 2021. [Online]. Available: https://arxiv.org/abs/2101.01466

[28] P. C. Parks, "A. M. Lyapunov's stability theory—100 years on," *IMA journal of Mathematical Control and Information*, vol. 9, no. 4, pp. 275–303, 1992.

[29] A. Forsgren, P. E. Gill, and M. H. Wright, "Interior methods for nonlinear optimization," *SIAM Rev.*, vol. 44, no. 4, pp. 525–597, 2002.

[30] P. T. Boggs and J. W. Tolle, "Sequential Quadratic Programming," *Acta Numer.*, vol. 4, no. 1995, pp. 1–51, 1995.

[31] H. K. Khalil, "Nonlinear systems third edition," *Patience Hall*, 2002.

[32] 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.

[33] K. H. Johansson and J. L. R. Nunes, "The Quadruple-Tank Process: A Multivariable Laboratory Process with an Adjustable Zero," *Proc. Am. Control Conf.*, vol. 8, no. 3, pp. 456–465, may 2000.