European Journal of Control 71 (2023) 100804

Contents lists available at ScienceDirect

European Journal of Control

journal homepage: www.elsevier.com/locate/ejcon

# Quickest physical watermarking-based detection of measurement replacement attacks in networked control systems[☆]

Arunava Naha[a,*], André Teixeira[b], Anders Ahlén[a], Subhrakanti Dey[a]

[a] Electrical Engineering, Uppsala University, Uppsala, Sweden
[b] Department of Information Technology, Uppsala University, Uppsala, Sweden

ARTICLE INFO

*Article history:*
Received 27 October 2022
Revised 28 January 2023
Accepted 19 March 2023
Available online 24 March 2023

Recommended by Prof. T Parisini

*Keywords:*
CUSUM test
Cyber-physical system
Deception attack
Linear quadratic gaussian control
Networked control system
Physical watermarking

ABSTRACT

In this paper, we propose and analyze an attack detection scheme for securing the physical layer of a networked control system (NCS) with a wireless sensor network against attacks where the adversary replaces the true observations with stationary false data. An independent and identically distributed watermarking signal is added to the optimal linear quadratic Gaussian (LQG) control inputs, and a cumulative sum (CUSUM) test is carried out using the joint distribution of the innovation signal and the watermarking signal for quickest attack detection. We derive the expressions of the supremum of the average detection delay (SADD) for a multi-input and multi-output (MIMO) system under the optimal and sub-optimal CUSUM tests. The SADD is asymptotically inversely proportional to the expected Kullback–Leibler divergence (KLD) under certain conditions. The expressions for the MIMO case are simplified for multi-input and single-output systems and explored further to distil design insights. We provide insights into the design of an optimal watermarking signal to maximize KLD for a given fixed increase in LQG control cost when there is no attack. Furthermore, we investigate how the attacker and the control system designer can accomplish their respective objectives by changing the relative power of the attack signal and the watermarking signal. Simulations and numerical studies are carried out to validate the theoretical results.

## 1. Introduction

Large distributed networked control systems (NCS) are getting deployed in various sectors such as manufacturing units, transportation systems, power systems, robotics, etc. [38]. Such cyber-physical systems (CPS) consist of embedded software, processors, wireless network, and other physical components. Along with their innumerable advantages, there is an increasing concern regarding safety and security. In the past, there have been several incidents of attack on CPS, *e.g.*, the Stuxnet attack [22], the attack on the sewage systems in Australia [1], the attack on the Davis-Besse nuclear power plant in Ohio, USA [5]. Attacks on such systems can cause loss of production, financial loss, a threat to human safety, etc. Securing CPS is a great challenge and it relies on both information security measures and system-theoretic approaches [35]. In general, the cyber layer is secured by employing various information security measures such as cryptography and digital watermarking to ensure the secure transmission and trustworthiness of data. However, information security measures alone cannot ensure the safety of the physical layer of the system, and they need to be complemented with system-theoretic approaches [35], for example, with such an approach studied in this manuscript.

There are two different attack strategies such as deception attack and denial of service (DoS) attack that adversaries usually apply to attack the physical layer of CPS [28,36]. In the DoS attack, the attacker makes the data unavailable possibly by jamming the wireless network [37]. On the other hand, under the deception attack, the adversary feeds the NCS with false data either by replacing or distorting the true observations and/or the control inputs [28,38]. In one scenario, the attacker records the true observations for a while and feeds the system with the recorded data along with some harmful exogenous inputs to remain stealthy. Such an attack strategy is called a replay attack [28]. In another class of deception attacks, also known as additive attacks or false data injection attacks, attackers add attack signals to the true observations or control signals [29,52].

[☆] This work is supported by The Swedish Research Council under grants 2017-04053 and 2018-04396, and by the Swedish Foundation for Strategic Research.

[*] Corresponding author.
*E-mail address:* arunava.naha@angstrom.uu.se (A. Naha).

https://doi.org/10.1016/j.ejcon.2023.100804
0947-3580/© 2023 The Author(s). Published by Elsevier Ltd on behalf of European Control Association. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

In this paper, we study a class of deception attacks where the attacker replaces the sensor measurements with fake observations which are statistically similar to the true measurements, which we call measurement replacement attacks in general. A well-established methodology to achieve this is to transmit the fake observations with a significantly stronger power than the sensor measurements. Therefore, the receiver in the wireless control system decodes the fake data as opposed to the true measurements since the received signal-to-interference and noise ratio is much larger for the fake measurements. Such measurement replacement attacks are also known as sensor spoofing attacks [25,50,51], often prevalent in GPS systems. In all the attack strategies, the attacker's objective is to make the system unstable or force the system to operate at a state outside its desired normal behaviour, and at the same time to remain stealthy as long as possible to cause maximum damage [28,37,38].

### 1.1. Related work

Several different approaches are found in the literature to secure CPS from attacks on the physical layer. In one approach, the security of the NCS is improved by designing attack resilient state estimators which can estimate the true states with bounded errors even if there is an attack [8,11,13]. In [7,33], the authors have studied different attack strategies which will be useful to design more resilient defence strategies. The defence strategies employed for attack detection can be broadly classified into two groups, *i.e.*, passive and active. In the passive attack detection scheme, the innovation signal is normally used as a residue signal with different statistical tests to detect attacks [14,30,34]. The passive detection schemes, in general, have an unsatisfactory probability of detection in the presence of noise and uncertainties. To address this problem, a data-driven residual generator is designed, and actuator attacks are detected by solving an $H_2/H_\infty$ mixed optimization problem in [24].

On the other hand, active attack detection schemes add physical watermarking signals to the control inputs to improve the probability of detection at the expense of an increased control cost [10,19,26–28,38]. Our paper follows this approach to design a detection scheme for measurement replacement-type deception attacks. The idea of physical watermarking is analogous to digital watermarking, which is used to authenticate the actual owner of digital content. In [27], the process of detecting a replay attack by adding a random Gaussian and independent and identically distributed (iid) watermarking signal to the linear quadratic Gaussian (LQG) control inputs is introduced. The statistics of the innovation signal changes in the presence of an attack, which is detected by a properly designed $\chi^2$ detector. The method is then improved in [26] by optimizing the watermarking signal power, and in [28] by generalizing the method and find the optimum watermarking signal in the class of Gaussian stationary processes. The $\chi^2$ based detection scheme is also studied for continuous-time systems in [48]. In the literature, there are attack detection schemes other than $\chi^2$, such as in [38], authors design two residue signals whose time averages will converge to some finite values when the system is under attack and demonstrate the method under a laboratory setup in [19]. The method is further generalized by considering the system model with a non-Gaussian process and observation noise in [39]. Stealthy false data injection attacks are detected using physical watermarking, and a non-linear auxiliary system in [15]. The problem of false data injection attacks in the presence of packet drop is studied in [47] by the design of a joint Bernoulli-Gaussian watermarking. In [49], the authors studied the design of completely stealthy FDI attacks and the necessary design conditions. The authors have also shown how an attacker can estimate the required parameters to launch such an attack from the input-output data. In [10], the authors reduce the increase of control cost by designing a periodic watermarking signal. In the context of sensor spoofing attacks, a secure trajectory planning problem is studied in [25], where the true signal from the Global Navigation Satellite System (GNSS) is replaced by the false position data from the attacker's system. In [23] a measurement replacement type stealthy attack mechanism on remote state estimation is studied. Most of the reported methods in the literature apply detection methods based on a large window of data samples and do not specifically address the problem of sequential quickest detection of attacks in CPS.

In this paper, we have studied the problem of quickest sequential attack detection. The research on quickest change detection can be traced back several decades [40]. We have taken the non-Bayesian approach of quickest change point detection where the change point or the attack point is unknown but deterministic as studied extensively in [17,21,42–44]. In [44], it is assumed that the data before and after the change point is iid. However, in our problem, the test data does not remain iid after the attack but is assumed to be stationary. To facilitate our analysis, we adopt the results in [17,21,42,43], which show that, under certain conditions, the cumulative sum (CUSUM) test provides the quickest change detection, *i.e.*, it minimises the supremum of the average detection delay (SADD) for a fixed upper limit on the average run length (ARL) to a false alarm for the general non-iid case. Since it is uncertain how long the system will be operational (especially in the case of an attack), the probability of false alarm (PFA) may not be a practically useful metric in this scenario [16,46]. Furthermore, the SADD asymptotically converges to the inverse of the expected value of the Kullback—;Leibler divergence (KLD) for the non-iid case, provided certain conditions are satisfied [17]. Note that KLD is an important and widely used measure of the disparity of two distributions, *i.e.*, post- and pre-attack distributions of the test data for our study. In our subsequent analysis, we refer to the CUSUM test using the conditional distribution for the non-iid case as the optimal CUSUM test. If the CUSUM test is performed using the unconditional distributions for the non-iid data, then we mention it as a sub-optimal CUSUM test. The latter may be applicable when the analytic form of the conditional distributions may be intractable.

### 1.2. Motivations and contributions

For the safety and security of CPS, it is of paramount importance to detect the attack with minimum possible delay to minimize the damage, thus favouring quickest sequential detection based methods. The watermarking based detection techniques reported in [10,28,38] are not specifically designed for the quickest detection of attacks. Thus we will here focus on the design and analysis of the quickest sequential detection of measurement replacement-type deception attacks, similar to Li and Ye [23], Liu et al. [25], Yılmaz and Arslan [50], Zhang et al. [51], by applying watermarking to the control inputs while keeping the system performance within a prescribed safety limit as recommended by the resilience requirements of CPS under attacks [6]. We consider a linear NCS where the attacker can hijack the sensor nodes and feed a stationary random process as fake measurement data to the estimator. The time of the attack is unknown but deterministic. The plant is controlled by a LQG controller, which receives the estimated states from a Kalman filter (KF). The controller adds a stationary but iid watermarking signal to the optimal control inputs and performs a CUSUM based test on the joint distribution of the innovation signal and the watermarking signal for the attack detection. We reported a preliminary study on this method for the scalar case applying sub-optimal CUSUM test, in [37]. In the current paper, we extend the work in [37] significantly by considering

**Table 1**
Notations.

| Symbol | Description |
|---|---|
| $\mathbb{R}^n$ | The set of $n \times 1$ real vectors |
| $\mathbb{R}^{m\times n}$ | The set of $m \times n$ real matrices |
| $\mathbf{A}^T$ | Transpose of matrix or vector $\mathbf{A}$ |
| $N(\mu, \boldsymbol{\Sigma})$ | Gaussian distribution with mean $\mu$ and variance $\boldsymbol{\Sigma}$ |
| $\{\cdot\} \cup \{\cdot\}$ | Union of two sets |
| $\boldsymbol{\Sigma} \geq \mathbf{0}$ | $\boldsymbol{\Sigma}$ is positive semi-definite matrix |
| $\boldsymbol{\Sigma} > \mathbf{0}$ | $\boldsymbol{\Sigma}$ is positive definite matrix |
| $\mathbf{x}_{a,k}$, $\mathbf{u}_{n,k}$, etc. | $k$th instant value of the corresponding variable |
| $[\cdot]_{ij}$ | $i$th row and $j$th column element of a matrix |
| $\lambda_{\gamma,i}$, $\lambda_{e,i}$, etc. | $i$th element of the corresponding vector |
| $\lvert \cdot \rvert$ | Determinant of a matrix or absolute value of a scalar |
| $tr(\cdot)$ | Trace of a matrix |
| $\{\mathbf{X}\}_1^{k-1}$ | $\{X_i : 1 \leq i \leq k-1\}$ |

more generalized system models, in-depth analysis of the optimal CUSUM test for the non-iid data, and provide extensive numerical simulations. Our main contributions are as follows. (i) We design a sequential quickest change detection test based on the CUSUM statistics that minimises the SADD subject to a lower bound on the ARL between two consecutive false alarms. To this end, we utilize the joint distribution of the innovation signal and the watermarking signal resulting in an increase in the KLD (and hence in a reduction in average detection delay), unlike some of the previous works which consider only the innovation signal. We also present a sub-optimal sequential detection technique that is useful where the optimal CUSUM test may not be tractable. (ii) We derive expressions of the expected KLD for the optimal CUSUM test and KLD for the sub-optimal case. An analysis of the behaviour of the KLD with respect to the watermarking signal power and attack signal power is performed, and some structural results are presented. (iii) We optimise the watermarking signal variance for a multi-input and single-output (MISO) system that maximises the expected KLD (optimal CUSUM test) or KLD (sub-optimal CUSUM test) subject to an upper bound on the increase in LQG control cost. Note: The proposed approach can be applied to detect a replay attack with a few modifications, as reported in [32]. Additionally, stealthy additive attacks [52] can also be detected with the proposed sequential attack detection mechanism by using an additional non-linear auxiliary system [15]. A detailed analysis of this approach is currently under investigation.

### 1.3. Paper organization

The organization of the remaining part of the paper is as follows. Section 2 describes the system model with the LQG controller and the attack strategy adopted for the paper. The mechanism of adding watermarking, the CUSUM test, and the associated detection delay are explained in Section 3. All the theorems and lemmas associated with multi-input and multi-output (MIMO) and MISO systems are provided in Section 4. The optimization technique to maximize the KLD by finding a proper watermarking signal variance is also illustrated in Section 4. We present numerical results in Section 5 to validate the theory. Section 6 concludes the paper.

### 1.4. Notations

We have used capital bold letters, *e.g.*, **A**, **B**, etc. to specify matrices and small bold letters, *e.g.*, **x**, **y**, etc. to specify vectors, unless specified otherwise. Some special notations are given in Table 1.

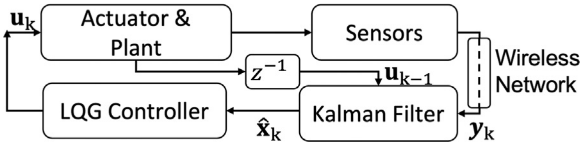

**Fig. 1.** Schematic diagram of the system during normal operation.

## 2. System and attack model

This section discusses the system model during the normal operations and under attack, and the attack strategy of the adversary considered in this paper.

### 2.1. System model during normal operations

We consider the following structure of the NCS, see Fig. 1 for a schematic diagram of the complete system during the normal operation,

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

Here $\mathbf{x}_k \in \mathbb{R}^n$ and $\mathbf{u}_k \in \mathbb{R}^p$ are the state and input vectors at the $k$th time instant respectively, whereas $\mathbf{w}_k \in \mathbb{R}^n \sim N(0, \mathbf{Q})$ is an iid process noise. $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times p}$, and $\mathbf{Q} \in \mathbb{R}^{n\times n}$. $\mathbf{Q} \geq \mathbf{0}$. Furthermore,

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

where $\mathbf{y}_k \in \mathbb{R}^m$ is the sensor output or the observation vector at the $k$th time instant. Here $\mathbf{C} \in \mathbb{R}^{m\times n}$, and $\mathbf{v}_k \in \mathbb{R}^m \sim N(0, \mathbf{R})$ is the iid measurement noise. We assume, $\mathbf{R} > \mathbf{0}$. The noise vectors $\mathbf{v}_k$ and $\mathbf{w}_k$ are mutually independent, and both are independent of the initial state vector, $\mathbf{x}_{k_0}$. The observations $\mathbf{y}_k$ are sent to the state estimator over a wireless network. We assume the system is stabilizable and detectable. We also assume that the system has been operational for a long time, thus the system is currently at steady state.

The Kalman filter (KF) uses the received sensor measurements and the input signal information, and estimates the states 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 $\Psi_k$ is the set of all measurements up to time $k$. The innovation $\gamma_k$ and steady state Kalman 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. $\mathbf{P}$ is the solution to 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 following infinite horizon LQG cost

$$J_c = \lim_{T\to\infty} E\left[\frac{1}{2T+1}\left\{\sum_{k=-T}^{T}\left(\mathbf{x}_k^T\mathbf{W}\mathbf{x}_k + \mathbf{u}_k^T\mathbf{U}\mathbf{u}_k\right)\right\}\right] \tag{8}$$

where $\mathbf{W} \in \mathbb{R}^{n\times n}$ and $\mathbf{U} \in \mathbb{R}^{p\times p}$ are positive definite diagonal weight matrices. The optimum input appears as a fixed gain linear control signal given by

$$\mathbf{u}_k^* = \mathbf{L}\hat{\mathbf{x}}_{k|k} \tag{9}$$

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

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

$$\mathbf{S} = \mathbf{A}^T\mathbf{S}\mathbf{A} + \mathbf{W} - \mathbf{A}^T\mathbf{S}\mathbf{B}\left(\mathbf{B}^T\mathbf{S}\mathbf{B} + \mathbf{U}\right)^{-1}\mathbf{B}^T\mathbf{S}\mathbf{A}. \tag{11}$$

### *2.2. Attack strategy and changes in system model*

The attack strategy of the adversary considered in this paper is discussed here. We assume that the attacker has the following knowledge about the system.

1. The attacker knows the system parameters **A**, **B**, **C**, **Q**, and **R**, and also the control policy, *i.e.*, **L**.
2. The attacker's system has sufficient energy to overpower the sensor nodes and make the wireless control receiver successfully decode the false measurements instead of the true sensor data.
3. The attacker does not have access to the control signal or the controller.

Note that in the information security literature, it is common to assume that the adversary has full knowledge of the system parameters and protocols according to the notion of "security without obscurity" also known as Kerckhoffs's principle, which essentially assumes that "the enemy knows the system" [41]. The attacker's knowledge of the system is a sensible assumption since the adversary can cause maximum damage under such a situation, which is essential to detect as fast as possible. However, even though we assume the adversary has complete knowledge of the system, to launch the measurement replacement attack, the attacker mainly uses the knowledge of the distribution of the true observation, as discussed below.

**Remark 1.** Assumption 3 has been incorporated because the control signal or controller is typically harder to access or manipulate than sensors, which are often more exposed and may be more easily accessible to an attacker. For instance, the measured plant may have moving parts, and sensors may not be directly attached to the plant, thus necessitating wireless sensing and communication of measurements. On the other hand, the control signal or the controller can be physically protected, for example, by being located inside a secure control room or only being accessible through secure communication channels. This assumption allows us to focus on the security of sensors and measurements transmitted to the controller. However, it is important to note that this assumption may not always be true, and protection against attacks on the control signal or controller should also be considered. Our attack detection methods can be adapted to address this, but it in not in the scope of the current paper, see Remark 3 for an example.

The objective of the adversary is to cause harm to the system by replacing the true sensor measurements $\mathbf{y}_k$ by fake observations $\mathbf{z}_k$, and at the same time remain stealthy. The adversary can achieve such a goal by overpowering the wireless sensor nodes. The measurement replacement type attacks studied in this paper are also called sensor spoofing attacks. Under the spoofing attack, the attacker sends the fake measurement signal to the receiver with a significantly higher power than he sensor nodes. As a result, the receiver accepts the fake measurements as legitimate while rejecting the true measurements from the sensor nodes [25,50].

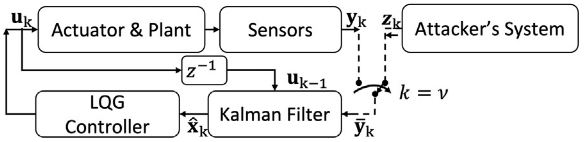

**Fig. 2.** Schematic diagram of the system under attack. $\bar{\mathbf{y}}_k = \mathbf{y}_k$ if $k < \nu, \bar{\mathbf{y}}_k = \mathbf{z}_k$ otherwise.

Figure 2 shows a schematic diagram of the system under attack. The system is assumed to be normal till the time $k < \nu$, and the attacker replaces the true observation $\mathbf{y}_k$ by the fake observation $\mathbf{z}_k$ at a deterministic but unknown time instant $k = \nu$, and keeps on sending the fake observation for $k \geq \nu$. A similar attack strategy is also studied in [23], where the stealthiness of the attack signal $\mathbf{z}_k$ is evaluated in terms of the KLD between the distributions of the fake and true observations as follows,

$$D(f_z, f_y) = \mathrm{E}_z\left[\frac{f_z(\mathbf{z_k})}{f_y(\mathbf{y_k})}\right]. \tag{12}$$

Here $f_z(\cdot)$ and $f_y(\cdot)$ denote the distributions of the fake and true observations, respectively. $\mathrm{E}_z[\cdot]$ means the expectation is taken with respect to the distribution of the fake data.

To summarize, the attacker's objective is to replace the true measurement $\mathbf{y}_k$ with fake data $\mathbf{z}_k$, which must appear statistically similar to $\mathbf{y}_k$ to remain stealthy, and at the same time to cause damage. In general, for linear control systems, the measurement vector $\mathbf{y}_k$ can be modelled as an autoregressive process. However, in this paper, we have adopted the following simple fake observation generator model using only a first-order Gauss-Markov autoregessive process to mimic the sensor measurements, which satisfies both the requirements of an attack signal, stealthiness and damage quality.

$$\mathbf{z}_k = \mathbf{A}_a\mathbf{z}_{k-1} + \mathbf{w}_{a,k-1}, \tag{13}$$

where $\mathbf{z}_k \in \mathbb{R}^m$, and $\mathbf{w}_{a,k} \sim N(0, \mathbf{Q}_a)$ is the iid noise vector at the $k$th time instant. $\mathbf{Q}_a \in \mathbb{R}^{m\times m}$ and $\mathbf{Q}_a \geq 0$. The stealthiness of the attack signal $\mathbf{z}_k$, *i.e.*, $D\left(f_z, f_y\right)$ (12), can be varied by selecting different values of $\mathbf{A}_a$ and $\mathbf{Q}_a$. In addition to that, since the true measurement $\mathbf{y}_k$ is stationary, the attacker will keep the fake measurement $\mathbf{z}_k$ stationary by taking the initial covariance of $\mathbf{z}_k$ as $\mathbf{E}_{zz}(0) \triangleq E\left[\mathbf{z_k}\mathbf{z_k^T}\right]$ to remain stealthy, where $\mathbf{E}_{zz}(0)$ is the solution to the following Lyapunov equation,

$$\mathbf{E}_{zz}(0) = \mathbf{A}_a\mathbf{E}_{zz}(0)\mathbf{A}_a^T + \mathbf{Q}_a. \tag{14}$$

The estimated states from the Kalman filter, $\hat{\mathbf{x}}^F$, take the following form when the system is under attack, *i.e.*, $k \geq \nu$,

$$\hat{\mathbf{x}}^F_{k|k-1} = \mathbf{A}\hat{\mathbf{x}}^F_{k-1|k-1} + \mathbf{B}\mathbf{u}_{k-1} \tag{15}$$

$$\hat{\mathbf{x}}^F_{k|k} = \hat{\mathbf{x}}^F_{k|k-1} + \mathbf{K}\widetilde{\gamma}_k \tag{16}$$

$$\widetilde{\gamma}_k = \mathbf{z}_k - \mathbf{C}\hat{\mathbf{x}}^F_{k|k-1}. \tag{17}$$

It is the same Kalman filter as given in (3)–(7) with the true observation $\mathbf{y}_k$ replaced by the fake data $\mathbf{z}_k$. So, the defender does not need to change anything for the Kalman filter during the attack.

The attacker can break open the feedback loop while masking the system trajectories from the Kalman filter by following the described attack model. Such a capability is particularly dangerous in systems that are open-loop unstable, as seen in the following example. For illustration, the true and estimated states, *i.e.*, $\mathbf{x}_k$ and $\hat{\mathbf{x}}_{k|k}$, respectively, of System-A, which is open-loop unstable, is plotted in Fig. 3, when the system is under attack from the time instant $k = 500$. See the model parameters of System-A from

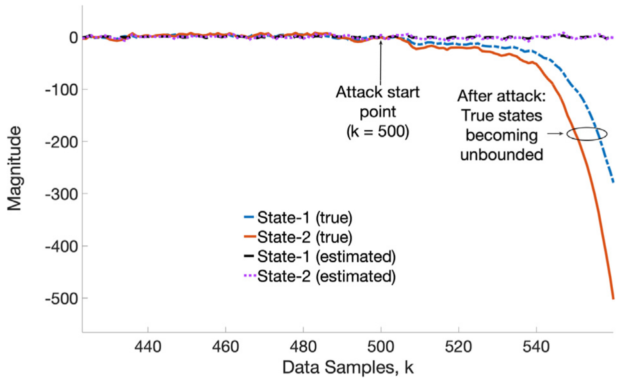

**Fig. 3.** True and estimated states of System-A.

Appendix H. The closed-loop system becomes unstable soon after the attack. In other words, Fig. 3 shows that the attack model in (13) can make the closed-loop system unstable for an open-loop unstable system, which may cause a significant amount of damage to CPS, thus illustrating the severity of such an attack model and demonstrating the need for the quickest attack detection mechanism to limit the damage.

**Remark 2.** If an attacker replaces only a subset of the true measurements with fake data, the attack signal can be modelled as,

$$\mathbf{z}_k^s = \mathbf{S}_a \mathbf{z}_k + (\mathbf{I}_n - \mathbf{S}_a)\mathbf{y}_k, \tag{18}$$

where $\mathbf{z}_k$ and $\mathbf{y}_k$ are given by (13) and (2), respectively. Here the *i*ith element of the diagonal matrix $\mathbf{S}_a$ will be 1 if the attacker chooses to replace the *i*th measurement, and 0 otherwise. It can be easily shown that under the attack model (18), the innovation signal still remains dependent on the watermarking signal, similar to the full measurement replacement attacks. Thus, we can apply Theorem 1 or Corollary 1.1 to detect the attack after evaluating the mean and variance of the after-attack innovation signal. However, a detailed discussion on such partial measurement replacement attacks is not included in the paper due to space constraints. Additionally, under replay attacks, the attacker commonly replaces all the observations with previously recorded data to mask the effect of the injection of exogenous signals to the true control inputs [28,38]. In [32], we have discussed how such replay attacks can be modelled by (13) and detected by the proposed sequential attack detection method.

**Remark 3.** As mentioned in the introduction, there are several other kinds of FDI attack models, such as stealthy data injection attacks studied in the literature [52]. A well-studied approach is to add an auxiliary system to the existing networked control system, which will not allow the attacker to remain stealthy [15]. Once the stealthiness is broken, we can apply the sequential detection mechanism studied in this paper for attack detection. A detailed discussion on such an attack detection mechanism is beyond the scope of the current article and is a topic of future research.

## 3. Physical watermarking based defence mechanism and delay in detection

This section proposes the physical-watermarking-based sequential attack detection scheme and discusses the delay in the detection process in different subsections. We perform the following hypothesis testing to detect the presence of the attack.

- $H_0$: No attack. Estimator receives the true observation $\mathbf{y}_k$
- $H_1$: Attack. Estimator receives a fake observation $\mathbf{z}_k$,

We design a sequential algorithm that minimizes the average delay (SADD) in detecting an attack (choosing $H_1$ when $H_1$ is true) subject to a lower bound on the ARL between false alarms (choosing $H_1$ when $H_0$ is true), see Section 3.4. We need the following information for the proposed detection scheme, the desired lower threshold of ARL, *i.e.*, $ARL_h$, the models of the test data generation process under $H_0$ and $H_1$, and the test data.

### 3.1. Selection of test data

The innovation signals ((17) and (5)) under attack and no attack contain different information. Therefore, the innovation signal is the natural selection of information source for hypothesis testing. The probability density functions (PDF) of $\gamma_k$ and $\tilde{\gamma}_k$ are denoted as $f_{\gamma_k}(\bar{\gamma}_k)$ and $f_{\tilde{\gamma}_k}(\bar{\gamma}_k)$ respectively, where $\bar{\gamma}_k = \gamma_k$ before attack, and $\bar{\gamma}_k = \tilde{\gamma}_k$ after attack. Both the distributions $f_{\gamma_k}(\bar{\gamma}_k)$ and $f_{\tilde{\gamma}_k}(\bar{\gamma}_k)$ are stationary in nature. The probability of attack detection will increase if the KLD i.e., $D\big(f_{\tilde{\gamma}_k}, f_{\gamma_k}\big)$, between the two distributions $f_{\tilde{\gamma}_k}(\bar{\gamma}_k)$ and $f_{\gamma_k}(\bar{\gamma}_k)$ under $H_1$ and $H_0$ increases [42],

$$D\big(f_{\tilde{\gamma}_k}, f_{\gamma_k}\big) = \int_{\mathbb{R}^m} f_{\tilde{\gamma}_k}(\bar{\gamma}) \log \frac{f_{\tilde{\gamma}_k}(\bar{\gamma})}{f_{\gamma_k}(\bar{\gamma})} d\bar{\gamma}. \tag{19}$$

The adversary will always try to remain stealthy by keeping the KLD low and thus cause maximum damage to the system. Therefore, the task of the control system designer is to maximize the KLD, thus making it difficult for the attacker to remain stealthy. Disturbances and measurement noise create uncertainty which favours the adversary.

### 3.2. Physical watermarking

A well-adopted technique to detect attacks on the control system is to add a watermarking signal, as described above [28,38]. The control designer thus adds a random watermarking signal $\mathbf{e}_k$ to the optimal LQG control input $\mathbf{u}_k^*$, see (20). The actual values of the watermarking signal will only be known to the controller and not to the attacker. However, the attacker may know the statistics of the watermarking signal.

$$\mathbf{u}_k = \mathbf{u}_k^* + \mathbf{e}_k \tag{20}$$

where $\mathbf{u}_k^*$ is the optimal input (9), $\mathbf{e}_k \sim N(0, \mathbf{\Sigma}_e)$ is an iid process, and $\mathbf{\Sigma}_e \geq 0$, and possibly non-diagonal matrix. In the literature, $\mathbf{e}_k$ is also taken to be a stationary Gauss-Markov process by some researchers. However, for our work, we assume it to be iid. The addition of $\mathbf{e}_k$ provides a means to the controller to check the authenticity of the measurement signal fed to the system. The distribution of the innovation signal will change substantially if the true measurement $\mathbf{y}_k$, which is correlated to $\mathbf{e}_{k-1}$, is replaced by $\mathbf{z}_k$, which is independent of $\mathbf{e}_{k-1}$, even if the attacker knows the statistics of $\mathbf{e}_k$.

Detection of the attack as early as possible is of utmost importance to reduce the damage. The optimal Neyman-Pearson (NP) test [28] and the asymptotic test [38] reported in the literature for the attack detection do not address the challenge of earliest detection. To this end, we have adopted a non-Bayesian sequential detection scheme [42] to detect the attack at the earliest time instant. It is assumed the attack takes place at a deterministic but unknown point in time. Instead of using the innovation signals $\gamma_k$ and $\tilde{\gamma}_k$ alone, we use the joint distributions of $\gamma_k$ and $\mathbf{e}_{k-1}$, and $\tilde{\gamma}_k$ and $\mathbf{e}_{k-1}$ for the test. We show the simulation results in the Section 5 that such a choice reduces the detection delay. The innovation signal during normal operation of the system and under attack will take the following forms (21) and (22), respectively,

$$\begin{aligned}\gamma_k &= \mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_{k|k-1} \\ &= \mathbf{CA}\big(\mathbf{x}_{k-1} - \hat{\mathbf{x}}_{k-1|k-1}\big) + \mathbf{Cw}_{k-1} + \mathbf{v}_k,\end{aligned} \tag{21}$$

$$
\begin{aligned}
\widetilde{\gamma}_k &= \mathbf{z}_k - \mathbf{C}\hat{\mathbf{x}}^F_{k|k-1} \\
&= \mathbf{z}_k - \mathbf{C}(\mathbf{A}+\mathbf{BL})\hat{\mathbf{x}}^F_{k-1|k-1} - \mathbf{CBe}_{k-1}.
\end{aligned} \tag{22}
$$

It is evident from (21) and (22) that the innovation signal during the normal operation of the system will be uncorrelated with the watermarking signal. However, on the contrary, the innovation signal will be correlated with the watermarking signal during the attack.

### 3.3. Detection delay

We use the average delay in the attack detection and the average run length to a false alarm as the metrics to measure the performance of the defence strategy. Here we adopt the theory of asymptotic optimality of the CUSUM test when the signal before and after the change (attack) may not be iid [42]. We start this section by introducing the definitions of relevant terms as follows.

**Average Detection Delay (ADD)**: ADD is defined as

$$ADD(\nu) \triangleq E_\nu[T_{H_1} - \nu | T_{H_1} > \nu] \tag{23}$$

where $E_\nu[\cdot]$ is the expectation taken with respect to the PDF under attack, when the attack start point is $\nu$. The attack start point $\nu$ is assumed to be unknown but deterministic in nature, whereas the random variable $T_{H_1}$ is the attack detection point in time or, in other words, the time instant of selecting Hypothesis $H_1$ by a hypothesis testing algorithm.

**Supremum Average Detection Delay (SADD)**: SADD is defined as

$$SADD \triangleq \sup_{1 \le \nu < \infty} E_\nu[T_{H_1} - \nu | T_{H_1} > \nu]. \tag{24}$$

**Average Run Length (ARL)**: ARL is defined as

$$ARL \triangleq E_\infty[T_{H_1}] \tag{25}$$

where $E_\infty[\cdot]$ is the expectation taken with respect to the PDF when there is no attack, *i.e.*, $\nu = \infty$. ARL represents the average time between two false alarms.

Ideally, we would like to have a detection scheme that will minimize ADD for any value of $\nu$ for a fixed threshold on ARL. However, such a detection scheme does not exist [42]. We can only find a procedure that will minimize the worst-case ADD over all $\nu$, *i.e.*, SADD, for a fixed threshold on ARL. As per the theory presented in [42], CUSUM is one such procedure. The CUSUM procedure is asymptotically minimax in the sense of minimizing the worst case ADD, *i.e.*, SADD, for all $\nu > 0$, as $ARL_h \to \infty$, and the minimum SADD is

$$SADD \sim \frac{\log(ARL_h)}{I} \tag{26}$$

where $I$ is a finite positive real number, $ARL_h$ is the threshold on ARL, $ARL \ge ARL_h$, provided the following three conditions are satisfied [42]:

$$\text{i) } \frac{1}{n}\lambda^{\nu}_{\nu+n} \xrightarrow[n\to\infty]{P_\nu} I, \tag{27}$$

$$
\begin{aligned}
&\text{ii) } \sup_{0\le\nu<\infty} \; ess\sup P_\nu \Big\{ M^{-1} \max_{0\le n<M} \lambda^{\nu}_{\nu+n} \ge \\
&(1+\epsilon) I | \Psi_\nu \Big\} \xrightarrow[M\to\infty]{} 0, \;\; \forall\, \epsilon > 0, \text{ and}
\end{aligned} \tag{28}
$$

$$
\begin{aligned}
&\text{iii) } \sup_{0\le\nu<k} \; ess\sup P_\nu \big\{ n^{-1}\lambda^{k}_{k+n} < I(1-\epsilon) | \Psi_\nu \big\} \xrightarrow[n\to\infty]{} 0, \\
&\forall\, 0 < \epsilon < 1 \;\; \text{and} \;\; k \ge 0
\end{aligned} \tag{29}
$$

where $P_\nu$ indicates the probability after the change and $M$ is a positive integer variable. Equation (27) implies that the left hand side will converge to $I$ in probability under $P_\nu$. Here $\Psi_\nu$ is the set of all observations up until the change point $\nu$. The variable $\lambda^{\nu}_{\nu+n}$ is defined as

$$\lambda^{\nu}_{\nu+n} \triangleq \sum_{k=\nu+1}^{n+\nu} \log \frac{f_{\nu,k}\big(X_k | \{\mathbf{X}\}_1^{k-1}\big)}{f_{\infty,k}\big(X_k | \{\mathbf{X}\}_1^{k-1}\big)} \tag{30}$$

where $X_k$ is the observation at the $k$th time instant and $\{\mathbf{X}\}_1^{k-1} = \{X_i : 1 \le i \le k-1\}$. In (30), $f_{\nu,k}(\cdot|\cdot)$ and $f_{\infty,k}(\cdot|\cdot)$ are the PDFs of the observations at the $k$th time instant for an attack starting at $\nu$ and without an attack, respectively.

For the case of attack detection using the joint distributions of innovation and watermarking signals,

$$\lambda^{\nu}_{\nu+n} = \sum_{k=\nu+1}^{n+\nu} \log \frac{f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}\big(\widetilde{\gamma}_k, \mathbf{e}_{k-1} | \{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-2}\big)}{f_{\gamma_k,\mathbf{e}_{k-1}}\big(\widetilde{\gamma}_k, \mathbf{e}_{k-1} | \{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-2}\big)} \tag{31}$$

where $f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}(\cdot|\cdot)$ and $f_{\gamma_k,\mathbf{e}_{k-1}}(\cdot|\cdot)$ are the joint conditional distributions of the innovation signal at the $k$th time instant and watermarking signal at $(k-1)$th time instant for the attack and no attack cases, respectively. $\{\bar{\gamma}\}_1^{k-1} = \{\gamma_i : 1 \le i < \nu\} \cup \{\widetilde{\gamma}_i : \nu \le i \le k-1\}$. The data ($\gamma_k$, $\widetilde{\gamma}_k$ and $\mathbf{e}_{k-1}$) satisfy the mean ergodicity theorem because of their stationarity property. The previously mentioned three conditions are satisfied under the mean ergodicity property of the data, and we can say $I$ converges to the expected value of the KLD between $f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}(\cdot|\cdot)$ and $f_{\gamma_k,\mathbf{e}_{k-1}}(\cdot|\cdot)$ as $n \to \infty$ [17]. In other words,

$$I \to \frac{1}{n} \sum_{k=\nu+1}^{n+\nu} \log \frac{f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}\big(\widetilde{\gamma}_k, \mathbf{e}_{k-1} | \{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-2}\big)}{f_{\gamma_k,\mathbf{e}_{k-1}}\big(\widetilde{\gamma}_k, \mathbf{e}_{k-1} | \{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-2}\big)},$$

as $n \to \infty$, which converges to the following form,

$$
\begin{aligned}
&E\Bigg[ \int_{\mathbb{R}^{m+p}} \log \frac{f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}\big(\widetilde{\gamma}_k, \mathbf{e}_{k-1} | \{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-2}\big)}{f_{\gamma_k,\mathbf{e}_{k-1}}\big(\widetilde{\gamma}_k, \mathbf{e}_{k-1} | \{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-2}\big)} \\
&f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}\big(\widetilde{\gamma}_k, \mathbf{e}_{k-1} | \{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-2}\big) d\gamma\, d\mathbf{e} \Bigg] \\
&= E\Big[ D\Big( f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}} | \{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-2} \Big) \Big].
\end{aligned} \tag{32}
$$

Here, the expectation is taken over the joint distribution of $\{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-2}$.

### 3.4. Optimal and sub-optimal CUSUM tests

The following CUSUM test will minimize the SADD asymptotically. The controller decides on hypothesis $H_0$ or $H_1$ based on the following test,

$$
\begin{aligned}
&H_0 : \text{Selected, when } gd_k < \log(ARL_h) \\
&H_1 : \text{Selected, when } gd_k \ge \log(ARL_h),
\end{aligned}
$$

where the CUSUM statistics $gd_k$ is evaluated as

$$
\begin{aligned}
&gd_k = \\
&\max\left(0, gd_{k-1} + \log \frac{f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}\big(\bar{\gamma}_k, \mathbf{e}_{k-1} | \{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-2}\big)}{f_{\gamma_{\mathbf{k}},\mathbf{e}_{k-1}}(\bar{\gamma}_k, \mathbf{e}_{k-1})}\right)
\end{aligned} \tag{33}
$$

where $\bar{\gamma}_k = \gamma_k$ before attack, and $\bar{\gamma}_k = \widetilde{\gamma}_k$ after attack, and

$$
\begin{aligned}
&SADD^* \to \frac{\log(ARL_h)}{E\Big[ D\Big( f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}} | \{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-2} \Big) \Big]}, \\
&\;\; as\ ARL_h \to \infty.
\end{aligned} \tag{34}
$$

Here, the threshold $ARL_h$ is a design parameter, which needs to be set by the system engineer from the detailed knowledge of the system and its vulnerabilities. A detailed study on the threshold selection for the CUSUM algorithm can be found in [20]. Since before the attack the innovation signal $\gamma_k$ and the watermarking signal $\mathbf{e}_{k-1}$ both are iids, and also independent to each other, the unconditional distribution is used in the denominator of (33). For certain cases, the closed-form expressions for the conditional distributions may not be found analytically, or it may be computationally too complex. Under such scenarios, the following sub-optimal CUSUM test can be carried out using the unconditional distributions for sequential attack detection,

$$g_k = \max\left(0, g_{k-1} + \log \frac{f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}(\bar{\gamma}_k, \mathbf{e}_{k-1})}{f_{\gamma_\mathbf{k},\mathbf{e}_{k-1}}(\bar{\gamma}_k, \mathbf{e}_{k-1})}\right). \tag{35}$$

Under the assumption that the system has been operating under a sufficiently long time, the joint distributions of the innovation and watermarking signal converge to their stationary distributions. Therefore, in what follows, we use only the stationary PDFs for the sub-optimal case. Under the sub-optimal CUSUM test, the SADD will converge as follows, since $I$ (26) converges to $D\left(f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}}\right)$.

$$SADD \to \frac{\log(ARL_h)}{D\left(f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}}\right)}, \quad as\ ARL_h \to \infty. \tag{36}$$

The test statistics $g_k$ is compared with the threshold $\log(ARL_h)$ as before.

### 3.5. Problem statement

The main problem to be addressed is threefold as follows.

1. The first task is to derive the expressions of the log-likelihood ratios, $\log \frac{f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}\left(\bar{\gamma}_k,\mathbf{e}_{k-1}|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-2}\right)}{f_{\gamma_\mathbf{k},\mathbf{e}_{k-1}}(\bar{\gamma}_k,\mathbf{e}_{k-1})}$ and $\log \frac{f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}(\bar{\gamma}_k,\mathbf{e}_{k-1})}{f_{\gamma_\mathbf{k},\mathbf{e}_{k-1}}(\bar{\gamma}_k,\mathbf{e}_{k-1})}$, needed to implement the optimal (33) and sub-optimal (35) CUSUM tests, respectively. The optimal CUSUM test solves the following optimization problem.

   $$\min_{T_{H_1}} SADD, \text{ s.t. } ARL \geq ARL_h.$$

2. The second task is to derive the analytical expressions of the KLDs, $E\left[D\left(f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}}|\{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-2}\right)\right]$ and $D\left(f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}}\right)$, under the optimal and sub-optimal CUSUM tests, respectively, to find asymptotic performance limits of SADD as $ARL \to \infty$, see (34) and (36).
3. The third task is to find an optimal watermarking covariance matrix $\Sigma_e$ that solves the following optimization problem,

   $$\min_{\Sigma_e} KLD, \text{ s.t. } \Delta LQG \leq J,$$

   where, $\Delta LQG$ is the increase in the LQG control cost due to the addition of the watermarking (in the case of no attack), and $J$ is a user defined threshold. Note that maximization of the KLD is equivalent to minimization of the asymptotic SADD (34).

## 4. Main results

We derive the expressions of the probability distributions, KLD and $\Delta LQG$ to evaluate the performance of the proposed detector analytically. We first state the results for the general MIMO systems in Section 4.1, and then simplify the theorems for the MISO systems in Section 4.2 to acquire better structural understanding. The technique to optimize the $\mathbf{\Sigma}_e$ to minimize SADD for a given upper bound on the $\Delta LQG$ is illustrated in Section 4.3.

### 4.1. Multiple input multiple output systems

We perform the following optimal CUSUM test to detect data deception attacks as stated in Theorem 1.

**Theorem 1.** *The optimal CUSUM test to detect the deception attack given by (13) will take the following form,*

$$gd_k = \max\left(0, gd_{k-1} + \log \frac{f_{\widetilde{\gamma}_k}\left(\bar{\gamma}_k|\{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-1}\right)}{f_{\gamma_\mathbf{k}}(\bar{\gamma}_k)}\right), \tag{37}$$

where $\bar{\gamma}_k = \gamma_k$ before attack, and $\bar{\gamma}_k = \widetilde{\gamma}_k$ after attack,

$$\begin{aligned}
&\left\{\widetilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-1}\right\} \\
&\sim N\left(\mu_{\widetilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}}, \mathbf{\Sigma}_{\widetilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}}\right), \\
&\mu_{\widetilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}} = \\
&\begin{cases} \mathbf{A_a z_{k-1}} - \mathbf{C}(\mathbf{A}+\mathbf{BL})\hat{\mathbf{x}}^\mathbf{F}_\mathbf{k-1|k-1} - \mathbf{CBe_{k-1}}, & k \geq \nu \\ \mathbf{A}_a\mathbf{y}_{k-1} - \mathbf{C}(\mathbf{A}+\mathbf{BL})\hat{\mathbf{x}}_{k-1|k-1} - \mathbf{CBe}_{k-1}, & k < \nu \end{cases}
\end{aligned} \tag{38}$$

$$\mathbf{\Sigma}_{\widetilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}} = \mathbf{Q}_a, \text{ and} \tag{39}$$

$$\begin{aligned}
&\gamma_k \sim N\left(\mathbf{0}, \mathbf{\Sigma}_\gamma\right), \\
&\mathbf{\Sigma}_\gamma = \mathbf{CPC}^T + \mathbf{R}.
\end{aligned} \tag{40}$$

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

Therefore, the optimal CUSUM test utilising the conditional distributions of the innovation signals before and after an attack is performed employing Theorem 1. The innovation signal $\gamma_k$ before an attack is iid, and independent of the watermarking signal $\mathbf{e}_{k-1}$. Therefore, the unconditional distribution is used in (37) for $\gamma_k$. On the other hand, the innovation signal after an attack $\widetilde{\gamma}_k$ is dependent on its previous values and watermarking signal values. Therefore, the conditional distribution of $\widetilde{\gamma}_k$ is used in (37), and the derived conditional mean and covariance are given in (38)-(39). The conditional variance is time-invariant. However, the conditional mean is changing for every time step depending on the previous measurement, estimated state and watermarking signal values.

**Remark 4.** The likelihood ratio in (37) will be evaluated using the innovation signal $\bar{\gamma}_k$ from the Kalman filter. $\bar{\gamma}_k = \gamma_k$ if $k < \nu$, and it will change automatically to $\bar{\gamma}_k = \widetilde{\gamma}_k$ if $k \geq \nu$ without any intervention from the defender. Similarly, $\mathbf{y}_k$ and $\hat{\mathbf{x}}_{k-1|k-1}$ will change to $\mathbf{z}_k$ and $\hat{\mathbf{x}}^\mathbf{F}_{k-1|k-1}$, respectively, after the attack, as given in (38). However, the attacker plays an active role by replacing the true observation $\mathbf{y}_k$ by the fake data $\mathbf{z}_k$ at $k \geq \nu$.

If we ignore the dependency of $\widetilde{\gamma}_k$ on its past values, we can simplify the CUSUM test as stated in Corollary 1.1. However, under such an assumption, the CUSUM test will not remain optimal.

**Corollary 1.1.** *The sub-optimal CUSUM test using the unconditional distributions to detect the deception attack given by (13) will take the following form,*

$$g_k = \max\left(0, g_{k-1} + \log \frac{f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}(\bar{\gamma}_k, \mathbf{e}_{k-1})}{f_{\gamma_\mathbf{k},\mathbf{e}_{k-1}}(\bar{\gamma}_k, \mathbf{e}_{k-1})}\right), \tag{41}$$

$$\text{where } \bar{\gamma}_k = \gamma_k \text{ before attack, and } \bar{\gamma}_k = \widetilde{\gamma}_k \text{ after attack,} \tag{42}$$

$$\gamma_{e,k} = \left[\gamma_k^T, \mathbf{e}_{k-1}^T\right]^T \sim N\left(\mathbf{0}, \mathbf{\Sigma}_{\gamma_e}\right),$$

$$\text{where } \boldsymbol{\Sigma}_{\gamma_e} = \begin{bmatrix} \boldsymbol{\Sigma}_{\gamma} & \mathbf{0}_{m\times p} \\ \mathbf{0}_{p\times m} & \boldsymbol{\Sigma}_e \end{bmatrix}, \text{ and} \tag{42}$$

$$\widetilde{\gamma}_{e,k} = \left[\widetilde{\gamma}_k^T, \mathbf{e}_{k-1}^T\right]^T \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}_{\widetilde{\gamma}_e}\right),$$

$$\text{where } \boldsymbol{\Sigma}_{\widetilde{\gamma}_e} = \begin{bmatrix} \boldsymbol{\Sigma}_{\widetilde{\gamma}} & -\mathbf{C}\mathbf{B}\boldsymbol{\Sigma}_e \\ -\boldsymbol{\Sigma}_e\mathbf{B}^T\mathbf{C}^T & \boldsymbol{\Sigma}_e \end{bmatrix}. \tag{43}$$

**Proof.** The proof of Corollary 1.1 is provided in Appendix B. □

Therefore, for the sub-optimal CUSUM test, the unconditional and asymptotically stationary distributions of $\gamma_k$ and $\widetilde{\gamma}_k$ are used. Such a test can be applied when designing the optimal CUSUM test is intractable, *e.g.*, replay attack detection as discussed in [32]. Also, for the optimal CUSUM test, the conditional mean needs to be evaluated at every time step, which increases the computational complexity compared to the sub-optimal CUSUM test.

**Remark 5.** Both the test statistics $gd_k$ and $g_k$ will be close to zero during the normal operation, and they will gradually increase after the attack on average over time.

To perform the sub-optimal CUSUM test as stated in Corollary 1.1, we need to evaluate the value of the covariance matrix of the innovation signal after the attack, $\widetilde{\gamma}_k$. We can derive the value of $\boldsymbol{\Sigma}_{\widetilde{\gamma}}$ using the following lemma.

**Lemma 1.** *The covariance matrix $\boldsymbol{\Sigma}_{\widetilde{\gamma}}$ of the innovation signal $\widetilde{\gamma}$ after the attack will take the following form,*

$$\begin{aligned} \boldsymbol{\Sigma}_{\widetilde{\gamma}} = \mathbf{E}_{zz}(0) &- \mathbf{C}(\mathbf{A}+\mathbf{B}\mathbf{L})\mathbf{E}_{xz}(-1) \\ &- [\mathbf{C}(\mathbf{A}+\mathbf{B}\mathbf{L})\mathbf{E}_{xz}(-1)]^T + \mathbf{C}\mathbf{B}\boldsymbol{\Sigma}_e\mathbf{B}^T\mathbf{C}^T \\ &+ \mathbf{C}(\mathbf{A}+\mathbf{B}\mathbf{L})\boldsymbol{\Sigma}_{x^F z}(\mathbf{A}+\mathbf{B}\mathbf{L})^T\mathbf{C}^T \\ &+ \mathbf{C}(\mathbf{A}+\mathbf{B}\mathbf{L})\boldsymbol{\Sigma}_{x^F e}(\mathbf{A}+\mathbf{B}\mathbf{L})^T\mathbf{C}^T, \end{aligned} \tag{44}$$

$$\text{where } \mathbf{E}_{xz}(-1) = \sum_{i=0}^{\infty} \mathcal{A}^i\mathbf{K}\mathbf{A}_a^{i+1}\mathbf{E}_{zz}(0) \tag{45}$$

*and* $\mathbf{E}_{zz}(0) = E\left[\mathbf{z}_k\mathbf{z}_k^T\right]$. $\boldsymbol{\Sigma}_{x^F z}$ *and* $\boldsymbol{\Sigma}_{x^F e}$ *are the solutions to the following Lyapunov equations,*

$$\begin{aligned} &\mathcal{A}\boldsymbol{\Sigma}_{x^F z}\mathcal{A}^T - \boldsymbol{\Sigma}_{x^F z} + \mathbf{K}\mathbf{E}_{zz}(0)\mathbf{K}^T + \mathcal{A}\mathbf{E}_{xz}(-1)\mathbf{K}^T \\ &\quad + \left(\mathcal{A}\mathbf{E}_{xz}(-1)\mathbf{K}^T\right)^T = 0, \text{ and} \end{aligned} \tag{46}$$

$$\mathcal{A}\boldsymbol{\Sigma}_{x^F e}\mathcal{A}^T - \boldsymbol{\Sigma}_{x^F e} + (\mathbf{I}_n - \mathbf{K}\mathbf{C})\mathbf{B}\boldsymbol{\Sigma}_e\mathbf{B}^T(\mathbf{I}_n - \mathbf{K}\mathbf{C})^T = 0. \tag{47}$$

*Here* $\mathcal{A} = (\mathbf{I}_n - \mathbf{K}\mathbf{C})(\mathbf{A}+\mathbf{B}\mathbf{L})$, *which is assumed to be strictly stable.* $\mathbf{I}_n$ *is a identity matrix of size $n \times n$.*

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

The analytical formula to derive the value of the unconditional variance $\boldsymbol{\Sigma}_{\widetilde{\gamma}}$ of the innovation signal $\widetilde{\gamma}$ under an attack as provided in Lemma 1 shows that $\boldsymbol{\Sigma}_{\widetilde{\gamma}}$ depends on the attacker's signal power, the watermarking signal power, and a few other system parameters. Furthermore, in addition to the sub-optimal CUSUM test, $\boldsymbol{\Sigma}_{\widetilde{\gamma}}$ is also needed for the derivation of the SADD under the optimal and sub-optimal CUSUM tests.

**Remark 6.** Since $\mathcal{A}$ is assumed to be strictly stable, the Lyapunov equations of (46) and (47) will have unique solutions. If $\mathcal{A}$ and $\mathbf{A}_a$ are not diagonalizable, then $\mathbf{E}_{xz}(-1)$ can be evaluated numerically by taking a large number of terms for the summation of (45) until the rest of the terms become negligible.

**Remark 7.** In order to use the quickest detection scheme, the pre and post-change pdfs must be known, which implies a priori knowledge of $\mathbf{A}_a$ and $\mathbf{Q}_a$, and may be impractical in a realistic setting. In such a case, $\mathbf{A}_a$ and $\mathbf{Q}_a$ can be estimated simultaneously with the proposed detection scheme from the received output (true or fake). Such a parameter estimation scheme can operate before and after the attack. However, before the attack, the estimates of $\mathbf{A}_a$ and $\mathbf{Q}_a$ will represent the healthy plant model. The estimates obtained can be then used simultaneously in the sequential attack detection algorithm. In the following Section 4.2.2, we have discussed a simple scheme to estimate $\mathbf{A}_a$ and $\mathbf{Q}_a$ from the received observations for a MISO system. A more detailed analysis of simultaneous parameter estimation and sequential attack detection algorithms is beyond the scope of the current manuscript.

We can derive a closed-form formula for $\mathbf{E}_{xz}(-1)$ (45), which is used in Lemma 1, provided $\mathcal{A}$ and $\mathbf{A}_a$ are diagonalizable, as given in the following corollary.

**Corollary 1.2.** *With the assumption that $\mathcal{A}$ and $\mathbf{A}_a$ are diagonalizable, $\mathbf{E}_{xz}(-1)$ will take the following form*

$$\mathbf{E}_{xz}(-1) = \mathbf{U}_{\mathcal{A}}\mathbf{T}_a\mathbf{U}_a^{-1}\mathbf{A}_a\mathbf{E}_{zz}(0). \tag{48}$$

*Here $\mathbf{U}_{\mathcal{A}}$ is the eigenvector matrix of $\mathcal{A}$, see (49). $\boldsymbol{\Sigma}_{\mathcal{A}} = diag\left[\lambda_{\mathcal{A},1}\ \lambda_{\mathcal{A},2}\ \cdots\right]$ is the eigenvalue matrix of $\mathcal{A}$ with the eigenvalues on its main diagonal. $\mathbf{U}_a$ is the eigenvector matrix of $\mathbf{A}_a$, see (50). $\boldsymbol{\Sigma}_a = diag\left[\lambda_{a,1}\ \lambda_{a,2}\ \cdots\right]$ is the eigenvalue matrix of $\mathbf{A}_a$ with the eigenvalues on its main diagonal.*

$$\mathcal{A} = \mathbf{U}_{\mathcal{A}}\boldsymbol{\Sigma}_{\mathcal{A}}\mathbf{U}_{\mathcal{A}}^{-1}. \tag{49}$$

$$\mathbf{A}_a = \mathbf{U}_a\boldsymbol{\Sigma}_a\mathbf{U}_a^{-1}. \tag{50}$$

*The ijth element of the $\mathbf{T}_a$ matrix is as follows*

$$[\mathbf{T}_a]_{ij} = \frac{[\mathbf{T}]_{ij}}{1 - \lambda_{\mathcal{A},i}\lambda_{a,j}}, \tag{51}$$

$$\text{and } \mathbf{T} = \mathbf{U}_{\mathcal{A}}^{-1}\mathbf{K}\mathbf{U}_a. \tag{52}$$

**Proof.** Proof of Corollary 1.2 is provided in the Appendix D. □

We evaluate the performance of the proposed attack detection technique in terms of SADD, which is inversely proportional to the expected KLD under the optimal CUSUM test and inversely proportional to the KLD under the sub-optimal CUSUM test, see (34) and (36). The expected KLD under the optimal CUSUM test and the KLD under the sub-optimal CUSUM test can be derived using the following theorem.

**Theorem 2.** *The expected KLD under the optimal CUSUM test $\left(E\left[D\left(f_{\widetilde{\gamma}_k}, f_{\gamma_k}|\{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-1}\right)\right]\right)$, and the KLD under the sub-optimal CUSUM test $\left(D\left(f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}}\right)\right)$ will be as follows,*

$$\begin{aligned} &E\left[D\left(f_{\widetilde{\gamma}_k}, f_{\gamma_k}|\{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-1}\right)\right] \\ &= \frac{1}{2}\left\{tr\left(\boldsymbol{\Sigma}_{\gamma}^{-1}\boldsymbol{\Sigma}_{\widetilde{\gamma}}\right) - m - \log\frac{|\,\mathbf{Q}_a\,|}{|\,\boldsymbol{\Sigma}_{\gamma}\,|}\right\}, \text{ and} \end{aligned} \tag{53}$$

$$\begin{aligned} &D\left(f_{\widetilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}}\right) \\ &= \frac{1}{2}\left\{tr\left(\boldsymbol{\Sigma}_{\gamma}^{-1}\boldsymbol{\Sigma}_{\widetilde{\gamma}}\right) - m - \log\frac{|\,\boldsymbol{\Sigma}_{\widetilde{\gamma}} - \mathbf{C}\mathbf{B}\boldsymbol{\Sigma}_e\mathbf{B}^T\mathbf{C}^T\,|}{|\,\boldsymbol{\Sigma}_{\gamma}\,|}\right\}. \end{aligned} \tag{54}$$

**Proof.** The proof of Theorem 2 is provided in Appendix E. □

Theorem 2 implies that the expected KLD and the KLD under the optimal and sub-optimal test, respectively, are largely dependent on the unconditional variances of the innovation signals $\boldsymbol{\Sigma}_{\gamma}$

and $\mathbf{\Sigma}_{\tilde{\gamma}}$ before and after an attack. They also depend on a few system and noise parameters. Furthermore, by subtracting (54) from (53), we get the difference between the expected KLD and the KLD to be $\log \frac{|\mathbf{\Sigma}_{\tilde{\gamma}} - \mathbf{CB\Sigma}_e\mathbf{B}^T\mathbf{C}^T|}{|\mathbf{Q}_a|}$, which corresponds to the optimality gap between the optimal and sub-optimal CUSUM tests. From (A.2), exploiting suitable independence properties of the involved random processes, it can be shown that $\mathbf{\Sigma}_{\tilde{\gamma}} - \mathbf{CB\Sigma}_e\mathbf{B}^T\mathbf{C}^T \geq \mathbf{Q}_a$. By eigenvalue comparison of the positive semidefinite matrices $\mathbf{\Sigma}_{\tilde{\gamma}} - \mathbf{CB\Sigma}_e\mathbf{B}^T\mathbf{C}^T$ and $\mathbf{Q}_a$, we can say $| \mathbf{\Sigma}_{\tilde{\gamma}} - \mathbf{CB\Sigma}_e\mathbf{B}^T\mathbf{C}^T | \geq | \mathbf{Q}_a |$, which ensures the optimality gap is positive.

Instead of taking the joint distribution of the innovation signal and the watermarking signal, if the optimal CUSUM test is performed using the conditional distribution of the innovation signal only, then the expected KLD will take the form of (55), which can be derived following the similar steps given in the Appendix A and Appendix E. An investigation of the KLD expression reveals that the numerator can be described as negative conditional differential entropy, which increases with further conditioning with respect to the watermarking signal, and the denominator (due to the Gaussian property of the distribution of the innovations) can be described as the conditional variance which decreases with further conditioning, thus resulting in the optimal KLD being larger than the expected KLD in (53).

The increase in KLD results in quicker attack detection on average due to (26). A detailed mathematical proof of this argument has been omitted due to the space constraints.

$$E\left[D\left(f_{\tilde{\gamma}_k}, f_{\gamma_k} | \{\bar{\gamma}\}_1^{k-1}\right)\right] = \frac{1}{2}\Big\{tr\Big(\mathbf{\Sigma}_{\gamma}^{-1}\Big(\mathbf{\Sigma}_{\tilde{\gamma}} - \mathbf{E}_{\mu} - \mathbf{E}_{\mu}^T\Big)\Big) \\ -m - \log \frac{| \mathbf{\Sigma}_{\tilde{\gamma}_k | \{\bar{\gamma}\}_1^{k-1}} |}{| \mathbf{\Sigma}_{\gamma} |}\Big\}, \tag{55}$$

where $\mathbf{\Sigma}_{\tilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1}} = \mathbf{Q}_a + (\mathbf{A}_a\mathbf{C} - \mathbf{C}(\mathbf{A}+\mathbf{BL}))\mathbf{G}(\mathbf{A}_a\mathbf{C} - \mathbf{C}(\mathbf{A}+\mathbf{BL}))^T + \mathbf{CB\Sigma}_e\mathbf{B}^\mathbf{T}\mathbf{C}^T$,

$$\mathbf{G} = \sum_{i=2}^{k-1} (\mathbf{A}+\mathbf{BL})^{i-1}\mathbf{B\Sigma}_e\mathbf{B}^\mathbf{T}\left[(\mathbf{A}+\mathbf{BL})^{i-1}\right]^T, \text{ and} \tag{56}$$

$$\mathbf{E}_{\mu} = (\mathbf{A}_a\mathbf{C} - \mathbf{C}(\mathbf{A}+\mathbf{BL}))\sum_{j=1}^{k-1}\sum_{i=2}^{j+1}(\mathbf{A}+\mathbf{BL})^{i-1}\mathbf{KE}_{\gamma e}(j-i+1)\mathbf{B}^T \\ \times\left[(\mathbf{A}+\mathbf{BL})^{j-1}\right]^T[(\mathbf{A}_a\mathbf{C} - \mathbf{C}(\mathbf{A}+\mathbf{BL}))]^T + (\mathbf{A}_a - \mathbf{C}(\mathbf{A}+\mathbf{BL})\mathbf{K}) \\ \times\sum_{j=1}^{k-1}\mathbf{E}_{\gamma e}(j)\mathbf{B}^T\left[(\mathbf{A}+\mathbf{BL})^{j-1}\right]^T[(\mathbf{A}_a\mathbf{C} - \mathbf{C}(\mathbf{A}+\mathbf{BL}))]^T, \tag{57}$$

$$\mathbf{E}_{\gamma e}(j) = \begin{cases} -\mathbf{C}(\mathbf{A}+\mathbf{BL})A^{j-2}(\mathbf{I}_n - \mathbf{KC})\mathbf{B\Sigma}_e & \text{if } j > 1 \\ \mathbf{0} & \text{otherwise.}\end{cases} \tag{58}$$

From the SADD expression in (34) and the expected KLD expression given in Theorem 2, we can say that SADD will be a finite quantity even without the watermarking, provided $\mathbf{\Sigma}_{\tilde{\gamma}} \neq \mathbf{\Sigma}_{\gamma}$. In other words, attacks can be detected without watermarking, but the detection delay will increase since the expected KLD reduces as $\mathbf{\Sigma}_e \to 0$. In the numerical results section, we have shown how SADD increases as $\Delta LQG \to 0$, which is equivalent to $\mathbf{\Sigma}_e \to 0$. As discussed before, the addition of watermarking also increases the control cost. The increase in the control cost during the normal system operation for the system model and the watermarking scheme considered in this paper is quantified in the following theorem.

**Theorem 3.** *The increase in the LQG cost ($\Delta LQG$) over the optimal LQG cost, when there is no attack, due to the addition of the watermarking signal is related to the watermarking signal covariance matrix $\mathbf{\Sigma}_e$ as follows,*

$$\Delta LQG = tr(\mathbf{H\Sigma}_e) \tag{59}$$

*where* $\mathbf{H} = \mathbf{B}^T\mathbf{\Sigma}_L\mathbf{B} + \mathbf{U}$ (60)

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

$$(\mathbf{A}+\mathbf{BL})^T\mathbf{\Sigma}_L(\mathbf{A}+\mathbf{BL}) - \mathbf{\Sigma}_L + \mathbf{L}^T\mathbf{UL} + \mathbf{W} = 0. \tag{61}$$

**Proof.** The theorem can be proved easily using Theorem 2 from Mo et al. [28], considering the iid watermarking as a special case of the hidden Markov model (HMM). □

**Remark 8.** Since the closed loop system $(\mathbf{A}+\mathbf{BL})$ is stable, the Lyapunov equation of (61) will have a unique solution.

Theorem 3 indicates the increase in the LQG control cost due to the addition of the watermarking, *i.e.*, $\Delta LQG$ is a linear function of the elements of the covariance matrix $\mathbf{\Sigma}_e$ of the added watermarking. The matrix $\mathbf{H}$ in (59) is dependent on the plant and controller parameters. Since the plant and the controller are assumed to be time-invariant, $\mathbf{H}$ will be a constant matrix during the steady-state operation of the system. Therefore, the increase in the LQG control cost is linear with respect to the covariance matrix, $\mathbf{\Sigma}_e$, of the watermarking signal.

### *4.2. Multiple input single output systems*

In this subsection, a simplified case of the MIMO system, *i.e.*, the MISO system is studied to get better structural understanding and insights. Lemma 2 provides the expressions for the expected KLD and KLD under the optimal and sub-optimal CUSUM tests, respectively, which are the simplified version of the KLD expressions provided in Theorem 2. The following attack model is assumed for the MISO system, which is a special case of the stochastic linear attack model given in (13),

$$E\left[z_k^2\right] = \sigma_z^2, \text{ and} \\ E\left[z_k z_{k-k_0}\right] = \rho^{k_0}\sigma_z^2,\ \rho < 1. \tag{62}$$

Therefore, $\mathbf{A}_a = \rho$, and $\mathbf{Q}_a = \left(1-\rho^2\right)\sigma_z^2$.

**Lemma 2.** *For a MISO system, the expected KLD $E\left[D\left(f_{\tilde{\gamma}_k}, f_{\gamma_k} | \{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-1}\right)\right]$ under the optimal CUSUM test, and the KLD $D\left(f_{\tilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}}\right)$ under the sub-optimal CUSUM test will be as follows,*

$$E\left[D\left(f_{\tilde{\gamma}_k}, f_{\gamma_k} | \{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-1}\right)\right] \\ = \frac{1}{2}\left\{\frac{\sigma_{\tilde{\gamma}}^2}{\sigma_{\gamma}^2} - 1 - \log\frac{(1-\rho^2)\sigma_z^2}{\sigma_{\gamma}^2}\right\}, \text{ and} \tag{63}$$

$$D\left(f_{\tilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}}\right) = \\ \frac{1}{2}\left\{\frac{\sigma_{\tilde{\gamma}}^2}{\sigma_{\gamma}^2} - 1 - \log\frac{\sigma_{\tilde{\gamma}}^2 - \mathbf{CB\Sigma}_e\mathbf{B}^T\mathbf{C}^T}{\sigma_{\gamma}^2}\right\} \tag{64}$$

*where the attack model is given by (62). $\sigma_{\gamma}^2$ and $\sigma_{\tilde{\gamma}}^2$ are the scalar variances of the innovation signals $\gamma_k$ and $\tilde{\gamma}_k$ before and after the attack, respectively. Hence,*

$$\sigma_{\gamma}^2 = \mathbf{CPC}^T + R, \text{ and} \tag{65}$$

$$\sigma_{\tilde{\gamma}}^2 = M_z\sigma_z^2 + \mathrm{tr}(\mathbf{M}_e\mathbf{\Sigma}_e) \tag{66}$$

*where $R$ and $M_z$ are scalar quantities. $M_z$ and $\mathbf{M}_e$ will take the following forms,*

$$M_z = 1 - 2\mathbf{C}(\mathbf{A}+\mathbf{BL})(\mathbf{I}_n - \rho A)^{-1}\mathbf{K}\rho+ \\ \mathbf{C}(\mathbf{A}+\mathbf{BL})\mathbf{\Sigma}_{x^F}^{z}(\mathbf{A}+\mathbf{BL})^T\mathbf{C}^T, \text{ and} \tag{67}$$

$$\mathbf{M}_e = \mathbf{B}^{\mathbf{T}}(\mathbf{I}_n - \mathbf{KC})^T \mathbf{\Sigma}^e_{x^F}(\mathbf{I}_n - \mathbf{KC})\mathbf{B} + \mathbf{B}^T\mathbf{C}^T\mathbf{CB} \tag{68}$$

*where $\mathbf{\Sigma}^z_{x^F}$ and $\mathbf{\Sigma}^e_{x^F}$ are the solutions to the Lyapunov equations,*

$$\begin{aligned} &A\mathbf{\Sigma}^z_{x^F}A^T - \mathbf{\Sigma}^z_{x^F} + \mathbf{KK}^T + A[\mathbf{I}_n - \rho A]^{-1}\mathbf{KK}^T\rho \\ &\quad + \left[A[\mathbf{I}_n - \rho A]^{-1}\mathbf{KK}^T\rho\right]^T = 0, \end{aligned} \tag{69}$$

and

$$A^T\mathbf{\Sigma}^e_{x^F}A - \mathbf{\Sigma}^e_{x^F} + (\mathbf{A} + \mathbf{BL})^T\mathbf{C}^T\mathbf{C}(\mathbf{A} + \mathbf{BL}) = 0$$

respectively. (70)

*Furthermore, $\Delta LQG$ coincides with Theorem 3.*

**Proof.** (63) and (64) can be derived directly by replacing the variables from (53) and (54) by their MISO system counterparts. Therefore, only the derivation of $\sigma^2_{\tilde{\gamma}}$ is provided in Appendix F. □

*4.2.1. Some structural results and their implications*

The expected KLD (63) and the KLD (64) from Lemma 2 are convex functions in $\sigma_z^2$. The convexity can be proved by taking the first and second derivative of (63) and (64) with respect to $\sigma_z^2$. The minimum value of the KLD will be achieved for $\sigma_z^{*2} = \frac{\sigma^2_{\tilde{\gamma}}}{M_z}$ and $\frac{\sigma^2_{\tilde{\gamma}} - tr((\mathbf{M}_e - \mathbf{B}^T\mathbf{C}^T\mathbf{CB})\mathbf{\Sigma}_e)}{M_z}$ for the optimal and sub-optimal tests, respectively. Therefore, we can conclude the KLD is not always increasing with the attacker signal power $\sigma_z^2$; it depends also on the power of the watermarking signal for the sub-optimal test. However, $\sigma_z^{*2}$ for the optimal test does not depend on the watermarking signal power. In fact, the attacker can modify $\sigma_z^2$ to $\sigma_z^{*2}$ to reduce the KLD which in turn reduces the probability of detection. On the other hand, the control system designer can choose $tr\left(\left(\mathbf{M}_e - \mathbf{B}^T\mathbf{C}^T\mathbf{CB}\right)\mathbf{\Sigma}_e\right) \geq \sigma^2_{\tilde{\gamma}}$ for the sub-optimal case, so that the KLD will always increase with the attacker signal power. However, under the optimal test, the control system designer can not do much to avoid this situation. On the other hand, for the sub-optimal test, the attacker needs to know $\Sigma_e$ to derive $\sigma_z^{*2}$.

*4.2.2. Estimation of $\mathbf{A}_a$ and $\mathbf{Q}_a$ for a MISO system*

As given in (62), for a MISO system, $\mathbf{A}_a = \frac{E[z_k z_{k-1}]}{\sigma_z^2}$, and $\mathbf{Q}_a = (1 - \mathbf{A}_a^2)\sigma_z^2$. Therefore, we need to estimate the variance $\sigma_z^2$ and the correlation $E[z_k z_{k-1}]$ from the received observations recursively. There are several recursive variance estimation algorithms available in the literature. However, we have used the following simple method [4],

$$\hat{\sigma}^2_{z,k} = B_z\hat{\sigma}^2_{z,k-1} + \frac{1 - B_z}{C_z}\mathbf{z}_k^2, \tag{71}$$

$$\hat{\sigma}^2_{zz,k} = B_z\hat{\sigma}^2_{z,k-1} + \frac{1 - B_z}{C_z}\mathbf{z}_k\mathbf{z}_{k-1}, \tag{72}$$

Here, $\hat{(\cdot)}$ denotes the estimated quantity. $\sigma^2_{zz,k} = E[z_k z_{k-1}]$. $0 < B_z < 1$, and $C_z$ is used to reduce the bias in the estimated quantity. We estimate $\mathbf{A}_a$ and $\mathbf{Q}_a$ from the beginning using the received observations $\mathbf{y}_k$ or $\mathbf{z}_k$ and use the estimates for the CUSUM test. However, as stated before, the estimates of $\mathbf{A}_a$ and $\mathbf{Q}_a$ will represent the healthy plant model before the attack. In Fig. 10, we have compared the ADD between the situations with known and estimated $\mathbf{A}_a$ and $\mathbf{Q}_a$ in the numerical results section.

*4.3. Optimum watermarking signal for MISO systems*

By increasing the watermarking power $\mathbf{\Sigma}_e$, we can improve the KLD, but at the same time, it also increases the control cost, *i.e.*, $\Delta LQG$ becomes larger. Therefore, we want to find the optimal $\mathbf{\Sigma}_e$, say $\mathbf{\Sigma}_e^*$, which will maximize the KLD subject to an upper bound on $\Delta LQG$. The optimization problem is formulated as follows,

$$\max_{\mathbf{\Sigma}_e} E\left[D\left(f_{\tilde{\gamma}_k}, f_{\gamma_k}|\{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-1}\right)\right]\text{or}$$

$$\max_{\mathbf{\Sigma}_e} D\left(f_{\tilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}}\right) \tag{73}$$

$$\text{s.t. } \Delta LQG \leq J \tag{74}$$

$$\mathbf{\Sigma}_e \geq 0 \tag{75}$$

where $J$ is a user choice. The positive semi-definite $\mathbf{\Sigma}_e$ matrix can be decomposed by the eigenvalue decomposition as

$$\mathbf{\Sigma}_e = \mathbf{V}_e\mathbf{\Lambda}_e\mathbf{V}_e^T, \tag{76}$$

where $\mathbf{V}_e$ is the orthonormal eigenvector matrix and $\mathbf{\Lambda}_e$ is the diagonal eigenvalue matrix. If we assume that $\mathbf{V}_e$ is known apriori, then we only need to find the optimum $\mathbf{\Lambda}_e$ which is a diagonal matrix. The optimization problem is simplified using the following theorem.

**Theorem 4.** *The optimum diagonal $\mathbf{\Lambda}_e$ that will maximize the expected KLD under the optimal CUSUM test or the KLD under the sub-optimal CUSUM test subject to $\Delta LQG \leq J$ will have only one non-zero element on its main diagonal.*

**Proof.** The proof of Theorem 4 is provided in Appendix G. □

In the light of Theorem 4, we search for the optimum $\mathbf{\Sigma}_e$ in the class of rank one positive semi-definitive matrices of the following form

$$\mathbf{\Sigma}_e = \lambda_e\mathbf{v}_e\mathbf{v}_e^T, \tag{77}$$

where $\lambda_e$ is the non-zero eigenvalue and $\mathbf{v}_e$ is the corresponding eigenvector. We modify (77) to represent it in the following form

$$\mathbf{\Sigma}_e = \mathbf{v}_\lambda\mathbf{v}_\lambda^T, \text{ where } \mathbf{v}_\lambda = \sqrt{\lambda_e}\mathbf{v}_e. \tag{78}$$

Finally, the optimization problem becomes,

$$\max_{\mathbf{v}_\lambda} E\left[D\left(f_{\tilde{\gamma}_k}, f_{\gamma_k}|\{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-1}\right)\right]\text{or}$$

$$\max_{\mathbf{v}_\lambda} D\left(f_{\tilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}}\right) \tag{79}$$

$$\text{s.t. } \Delta LQG \leq J. \tag{80}$$

The optimization problem can be solved using different methods such as the sequential quadratic programming (SQP) [2], the interior point method [12], etc. However, in this paper, we use a simple gradient descent based algorithm to solve the optimization problem (79)–(80) for a MISO system as follows.

For the optimal CUSUM test, using (53), (77), and (78), we can say that maximization of $E\left[D\left(f_{\tilde{\gamma}_k}, f_{\gamma_k}|\{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-1}\right)\right]$ with respect to $\mathbf{v}_\lambda$ is the same as maximizing the following function with respect to $\mathbf{v}_\lambda$.

$$\begin{aligned} &\mathbf{v}_\lambda^T\mathbf{H}_{KLD}\mathbf{v}_\lambda \\ &\text{where} \\ &\mathbf{H}_{KLD} = \mathbf{B}^T(\mathbf{I}_n - \mathbf{KC})^T L_e(\mathbf{I}_n - \mathbf{KC})\mathbf{B} + \mathbf{B}^T\mathbf{C}^T\mathbf{CB} \end{aligned} \tag{81}$$

and $L_e$ is the solution to the Lyapunov equation

$$A^T L_e A - L_e + (\mathbf{A} + \mathbf{BL})^T\mathbf{C}^T\mathbf{C}(\mathbf{A} + \mathbf{BL}) = 0 \tag{82}$$

Since the matrix $A$ is assumed to be strictly stable, the Lyapunov equation of (82) will have unique solution. (81) and (82) can be simplified further for the system with a relative degree higher than one, since $\mathbf{CB} = \mathbf{0}$. We take the negative of the cost function to convert the optimization problem into a minimization one. We then derive the Lagrangian, its first and second derivatives. Using (81) and (59) the Lagrangian can be written in the following form

$$L(\mathbf{v}_\lambda, \mu) = -\mathbf{v}_\lambda^T \mathbf{H}_{KLD} \mathbf{v}_\lambda + \mu\left(\mathbf{v}_\lambda^T \mathbf{H} \mathbf{v}_\lambda - J\right). \tag{83}$$

The first derivatives of $L(\mathbf{v}_\lambda, \mu)$ with respect to $\mathbf{v}_\lambda$ and $\mu$ take the following forms

$$\nabla_{\mathbf{v}_\lambda} L(\cdot) = \mathbf{C}_c \mathbf{v}_\lambda, \text{ and} \tag{84}$$

$$\frac{\partial}{\partial \mu} L(\cdot) = \mathbf{v}_\lambda^T \mathbf{H} \mathbf{v}_\lambda - J \tag{85}$$

$$\text{where } \mathbf{C}_c = -2\mathbf{H}_{KLD} + 2\mu \mathbf{H}. \tag{86}$$

The second derivatives of $L(\cdot)$ with respect to $\mathbf{v}_\lambda$, *i.e.*, the Hessian matrix $\mathbf{H}_s$, becomes

$$\mathbf{H}_s = \nabla^2_{\mathbf{v}_\lambda} L(.) = \mathbf{C}_c^T. \tag{87}$$

On the other hand, for the sub-optimal CUSUM test, we form the Lagrangian using (64), (68), and (59) for a MISO system as follows.

$$\begin{aligned} L(\mathbf{v}_\lambda, \mu) = &-\frac{1}{2}\left(\frac{M_z\sigma_z^2 + \mathbf{v}_\lambda^T \mathbf{M}_{ev} \mathbf{v}_\lambda + \mathbf{v}_\lambda^T \mathbf{B}^T \mathbf{C}^T \mathbf{C} \mathbf{B} \mathbf{v}_\lambda}{\sigma_\gamma^2}\right) \\ &- \frac{1}{2} + \frac{1}{2}\log\left(M_z\sigma_z^2 + \mathbf{v}_\lambda^T \mathbf{M}_{ev} \mathbf{v}_\lambda\right) - \frac{1}{2}\log\left(\sigma_\gamma^2\right) \\ &+ \mu\left(\mathbf{v}_\lambda^T \mathbf{H} \mathbf{v}_\lambda - J\right), \end{aligned} \tag{88}$$

where $\mathbf{M}_{ev}$ is the first part of the right hand side of (68), *i.e.*, $\mathbf{M}_{ev} = \mathbf{B}^{\mathbf{T}}(\mathbf{I}_n - \mathbf{KC})^T \boldsymbol{\Sigma}^e_{x^F}(\mathbf{I}_n - \mathbf{KC})\mathbf{B}$. The first derivatives of $L(\mathbf{v}_\lambda, \mu)$ with respect to $\mathbf{v}_\lambda$ and $\mu$ take the following forms,

$$\begin{aligned} \nabla_{\mathbf{v}_\lambda} L(.) = &-\frac{1}{\sigma_\gamma^2}\left(\mathbf{M}_{ev}\mathbf{v}_\lambda + \mathbf{B}^T \mathbf{C}^T \mathbf{C} \mathbf{B} \mathbf{v}_\lambda\right) \\ &+ \frac{\mathbf{M}_{ev}\mathbf{v}_\lambda}{M_z\sigma_z^2 + \mathbf{v}_\lambda^T \mathbf{M}_{ev} \mathbf{v}_\lambda} + 2\mu \mathbf{H} \mathbf{v}_\lambda = \mathbf{C}_c \mathbf{v}_\lambda, \text{ and} \end{aligned} \tag{89}$$

$$\frac{\partial}{\partial \mu} L(.) = \mathbf{v}_\lambda^T \mathbf{H} \mathbf{v}_\lambda - J, \tag{90}$$

$$\begin{aligned} &\text{where } \mathbf{C}_c = \mathbf{C}_{ca} + \mathbf{v}_\lambda^T \mathbf{M}_{ev} \mathbf{v}_\lambda \mathbf{C}_{cb}, \\ &\mathbf{C}_{ca} = \left(1 - \frac{\mathbf{M}_z\sigma_z^2}{\sigma_\gamma^2}\right)\mathbf{M}_{ev} - \frac{\mathbf{M}_z\sigma_z^2}{\sigma_\gamma^2}\mathbf{B}^T \mathbf{C}^T \mathbf{C} \mathbf{B} + 2\mu \mathbf{M}_z {\sigma_z}^2 \mathbf{H}, \\ &\text{and } \mathbf{C}_{cb} = 2\mu \mathbf{H} - \frac{1}{\sigma_\gamma^2}\mathbf{M}_{ev} - \frac{1}{\sigma_\gamma^2}\mathbf{B}^T \mathbf{C}^T \mathbf{C} \mathbf{B}. \end{aligned} \tag{91}$$

The second derivative of $L(\cdot)$ with respect to $\mathbf{v}_\lambda$, *i.e.*, the Hessian matrix $\mathbf{H}_s$, becomes

$$\mathbf{H}_s = \nabla^2_{\mathbf{v}_\lambda} L(.) = \mathbf{C}_{ca}^T + 2\mathbf{M}_{ev} \mathbf{v}_\lambda \mathbf{v}_\lambda^T \mathbf{C}_{cb}. \tag{92}$$

A primal-dual approach to find the optimum $\boldsymbol{\Sigma}_e$ is provided in Algorithm 1.

The step sizes ($s_k$, $K_{\mu,k}$) can be derived at every step using the backtracking algorithm [3], which ensures the convergence to some local optima since the Hessian matrices under both the tests are indefinite matrices. Next, we briefly discuss the computational runtime complexity of the proposed detection scheme.

**Algorithm 1** To find optimum $\boldsymbol{\Sigma}_e$.

```
Initialize: $s_0$, $K_{\mu,0}$, $max\_iteration$, and $\mu = 0$.
for $k = 1 : \max\_iteration$ do
    Find the best solution $\mathbf{v}^*_{temp}$ for the set of equations, $\nabla_{\mathbf{v}_\lambda} L(.) = 0$ and $\frac{\partial}{\partial\mu} L(.) = 0$.
    if $\mathbf{v}^{T*}_{temp} \mathbf{H} \mathbf{v}^*_{temp} - J \neq 0$ then
        $\mu \leftarrow \mu + s_k \frac{\partial}{\partial\mu} L(.)$
    else
        if $\mathbf{H}_s \geq 0$ then
            $\mathbf{v}^*_\lambda \leftarrow \mathbf{v}^*_{temp}$
            break
        else
            $\mu \leftarrow \mu + K_{\mu,k}\left(-\frac{\partial}{\partial\mu} L(.)\right)$
        end if
    end if
end for
$\boldsymbol{\Sigma}_e = \mathbf{v}^*_\lambda \left[\mathbf{v}^*_\lambda\right]^T$
```

### 4.4. Computational complexity

The proposed technique is an online method. At run time, we only need to evaluate $gd_k$ (37) and compare the test statistics with a fixed threshold at each time step. For our problem formulation, most of the heavy computations, such as matrix inversion and computation of determinants, associated with the evaluation of the likelihood ratio (37) can be derived offline since the variances are fixed, see (39)–(40). The most expensive operations at runtime are a few matrix-vector multiplications with the highest computational complexity of $O(np)$, see (38).

## 5. Numerical results

In this section, we will illustrate and validate different aspects of the theorems and lemmas presented in this paper using two different system models. The two different systems are (i) System-A: A second-order open-loop unstable MISO system, and (ii) System-B: A fourth-order open-loop stable MIMO system. The system parameters are provided in Appendix H. System-B is a linearized minimum phase quadruple tank system [18] which has been used previously to test a deception attack detection scheme in the literature [10]. The model provided in [18] is a continuous-time model, which is discretized with a sampling time of 2s, which is similar to Fang et al. [10]. In our work, only the level sensor gains have been are altered to make the magnitude of the product **CB** numerically significant.

### 5.1. Tradeoff between SADD and ΔLQG under optimal CUSUM test

Figure 4 shows the tradeoff between the SADD and the increase in the LQG control cost $\Delta LQG$ for System-A and System-B under the optimal CUSUM test (37). We plot the derived SADD using the theory developed in this paper, and the estimated SADD from Monte-Carlo (MC) simulation, where $\boldsymbol{\Sigma}_e$ is assumed to be diagonal and all the watermarking signals have equal power. An increase in LQG cost results in quicker detection.

### 5.2. Benefit of using the joint distribution

The choice of the joint distribution of the innovation signal and the watermarking signal improves the KLD for a fixed $\Delta LQG$ value compared to the case where the joint distribution is not considered. Therefore, we achieve the same SADD at a lower cost. As shown in Fig. 5, the same theoretical SADD can be achieved at

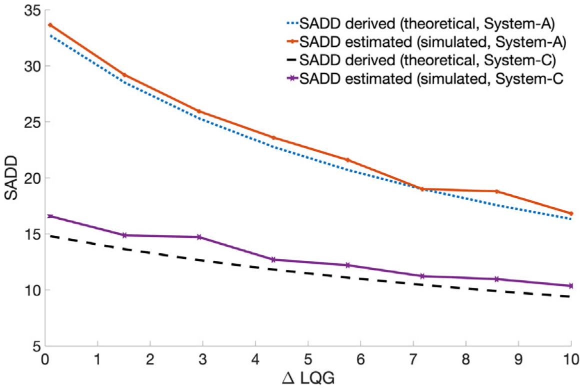

**Fig. 4.** SADD vs. $\Delta LQG$ plot for System-A and System-B.

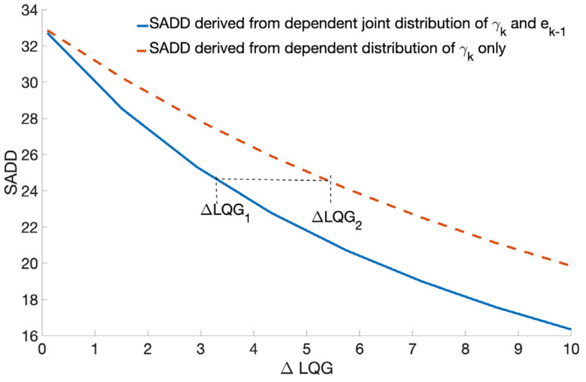

**Fig. 5.** Comparison of SADD vs. $\Delta LQG$ plots between the optimal CUSUM detection schemes using joint and single distributions for System-A.

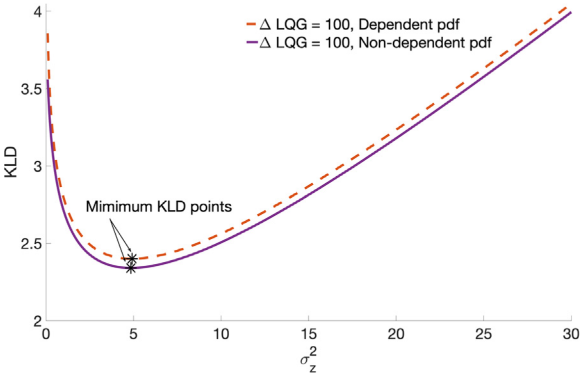

**Fig. 6.** KLD vs. $\sigma_z^2$ plots for System-A.

64% (approx.) reduced $\Delta LQG$ for System-A between the $\Delta LQG_1$ and $\Delta LQG_2$ points under the optimal CUSUM test. The percentage reduction in $\Delta LQG$ is evaluated as $\frac{\Delta LQG_2 - \Delta LQG_1}{\Delta LQG_1} \times 100\%$.

### 5.3. Convexity of KLD with respect to $\sigma_z^2$

Figure 6 shows how the KLD varies with $\sigma_z^2$ for System-A under the optimal and sub-optimal CUSUM tests. The KLD appears to be a convex function with respect to $\sigma_z^2$, and the minimum points are the same as predicted by our theory, see Fig. 6. We assume, $\Delta LQG = 100$, and $\boldsymbol{\Sigma}_e$ to be diagonal and both the watermarking signals to have equal power. Figure 6 can also be interpreted as, for the selected $\Delta LQG$ we can detect an attack equally well for a small $\sigma_z^2$ as for a significantly larger $\sigma_z^2$.

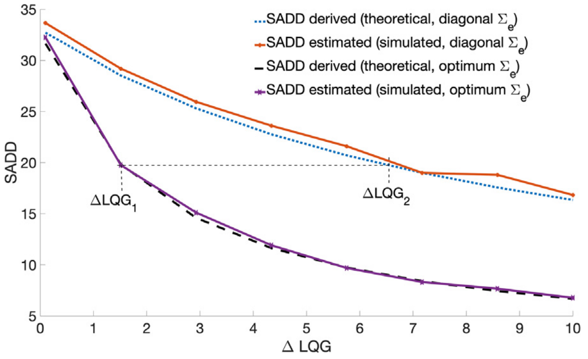

**Fig. 7.** SADD vs. $\Delta LQG$ plot for System-A with optimum and non-optimum $\boldsymbol{\Sigma}_e$ under optimal CUSUM test.

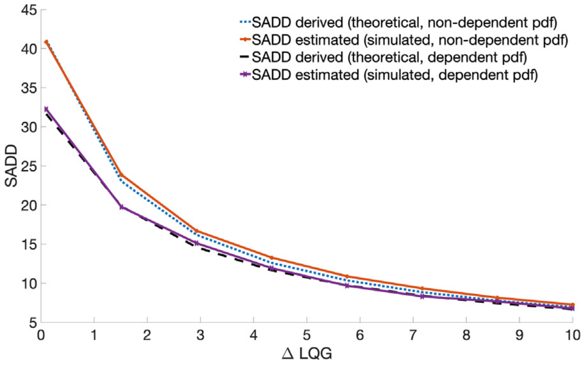

**Fig. 8.** SADD vs. $\Delta LQG$ plot for System-A under optimal and sub-optimal CUSUM tests.

### 5.4. Optimum vs non-optimum $\boldsymbol{\Sigma}_e$

We optimize the $\boldsymbol{\Sigma}_e$ under the optimal test using the method in Section 4.3. Figure 7 shows the SADD vs $\Delta LQG$ plots using the optimized $\boldsymbol{\Sigma}_e$ and a diagonal $\boldsymbol{\Sigma}_e$ with equal signal power under the optimal CUSUM test. We plot the derived SADD using our theory and the estimated SADD from MC simulation for optimized $\boldsymbol{\Sigma}_e$ and non-optimized $\boldsymbol{\Sigma}_e$ in the figure. It is evident that optimizing $\boldsymbol{\Sigma}_e$ helps in improving SADD for a fixed upper limit on $\Delta LQG$. On the other hand, we can comment that the same theoretical SADD can be achieved at 336% (approx.) reduced $\Delta LQG$ for System-A between the points $\Delta LQG_1$ and $\Delta LQG_2$.

### 5.5. Optimal vs sub-optimal CUSUM

Figure 8 illustrates the advantage of performing the optimal CUSUM test with conditional PDFs over the sub-optimal CUSUM test using the unconditional PDFs for System-A. For both the plots, optimum $\boldsymbol{\Sigma_e}$ has been used for the corresponding cases. Therefore, we can achieve lower SADD for the same $\Delta LQG$ with the optimal CUSUM test compared to the sub-optimal one. The benefit is larger for the lower $\Delta LQG$ values as per the figure.

### 5.6. Comparison between the proposed method with known and estimated $\mathbf{A}_a$ and $\mathbf{Q}_a$ and the optimal NP detector

We have compared the optimal CUSUM test results for known and estimated $\mathbf{A}_a$ and $\mathbf{Q}_a$ with the optimal NP detector based method reported in [26,28]. The watermarking signal is taken to be iid, and the $\boldsymbol{\Sigma}_e$ is optimized for both the cases. In [28], the optimal NP detector rejects the $H_0$ hypothesis in favour of $H_1$ if

$$g_{NP,k}(\gamma_k, \mathbf{e}_{k-1}, \cdots) = \gamma_k^T \boldsymbol{\Sigma}_\gamma^{-1} \gamma_k$$

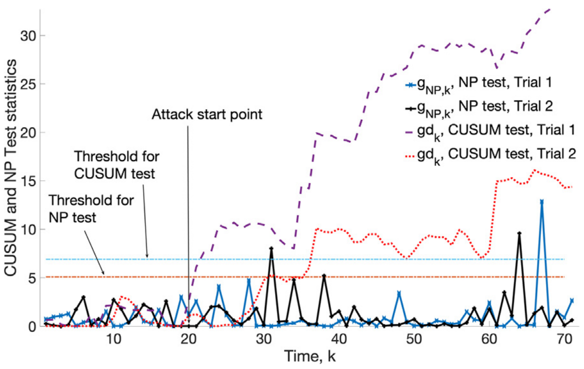

**Fig. 9.** Test statistics under optimal CUSUM test and optimal NP test.

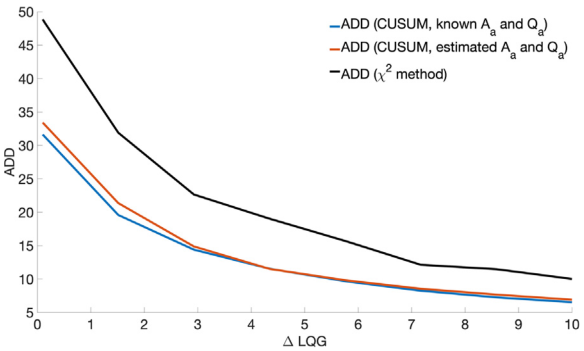

**Fig. 10.** ADD vs. $\Delta LQG$ plot for System-A under optimal CUSUM test with known and estimated $\mathbf{A}_a$ and $\mathbf{Q}_a$ and optimal NP test.

$$-\left(\gamma_k-\mu_{NP,k}\right)^T\left(\boldsymbol{\Sigma}_\gamma+\boldsymbol{\Sigma}_f\right)^{-1}\left(\gamma_k-\mu_{NP,k}\right)\geq\eta \tag{93}$$

where $\mu_{NP,k}=-\mathbf{C}\sum_{i=-\infty}^{k}A^{k-i}\mathbf{B}\mathbf{e}_i,$ (94)

$$\boldsymbol{\Sigma}_f=\mathbf{C}L_f\mathbf{C}^T, \text{ and} \tag{95}$$

$$L_f=AL_fA^T+\mathbf{B}\boldsymbol{\Sigma}_e\mathbf{B}^T. \tag{96}$$

The threshold $\eta$ is estimated by simulation from

$$P_\infty\{g_{NP,k}(\cdot)\geq\eta\}=\alpha \tag{97}$$

where $P_\infty$ denotes the probability under no attack condition, and $\alpha$ is the threshold on the false alarm rate. The false alarm rate is the reciprocal of the ARL [31,45]. For the method in [28], the ADD is estimated as

$$ADD_{NP}=E\left[\inf\{k:g_{NP,k}(\cdot)\geq\eta\}\right]. \tag{98}$$

Figure 9 illustrates how the test statistics $gd_k$ and $g_{NP,k}$ vary with time $k$ under the optimal CUSUM (37) with known $\mathbf{A}_a$ and $\mathbf{Q}_a$ and NP tests for two random trial runs. The thresholds for the corresponding tests are also shown in the figure. When the test statistics crosses the threshold for the first time that is considered as the attack detection point. System-A is used for generating Fig. 9. It should be noted that the CUSUM test minimizes the worst-case ADD, *i.e.*, SADD for a fixed lower limit on ARL, which means that although the detection delay may be higher for the CUSUM test compared to the NP detector for a specific random trial, on average, its detection delay, SADD, will always be lower than the NP detector.

Figure 10 shows the trade-off between the ADD and the increase in $\Delta LQG$ for System-A under the optimal CUSUM test with known and estimated $\mathbf{A}_a$ and $\mathbf{Q}_a$, and the method reported in [28]. We plot the estimated ADD applying the optimal CUSUM test on the simulated data for known and estimated $\mathbf{A}_a$ and $\mathbf{Q}_a$, and the estimated ADD applying the test reported in [28]. The attack start point was fixed at $\nu=500$. We observe a slightly higher ADD for the estimated $\mathbf{A}_a$ and $\mathbf{Q}_a$ case compared to the completely known situation. Notably, however, the proposed detection scheme with estimated $\mathbf{A}_a$ and $\mathbf{Q}_a$ performs better than the optimal NP detector. $ADD_{NP}$ is 61.5% (approx.) and 45.8% (approx.) higher than the ADD from the proposed method with known $\mathbf{A}_a$ and $\mathbf{Q}_a$ at $\Delta LQG=1.51$ and $\Delta LQG=7.17$, respectively. The difference is evaluated as $(ADD_{NP}-ADD)/ADD\%$.

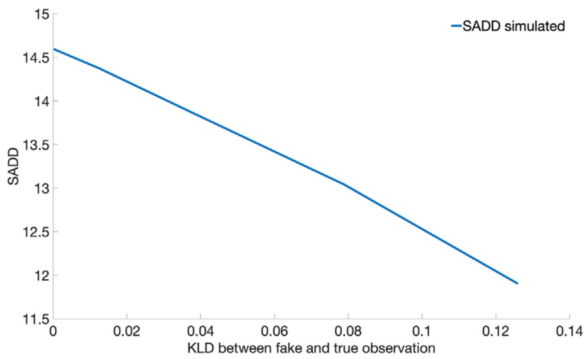

**Fig. 11.** SADD vs KLD between fake and true observations for System-A. $\Delta LQG=5$.

### *5.7. SADD vs stealthiness comparison*

In Fig. 11, we have plotted the SADD from the proposed detection scheme with a known attacker system model vs. different degrees of stealthiness of the attack signal using simulated data from System-A by varying $\mathbf{A}_a$ and $\mathbf{Q}_a$. As discussed before, the stealthiness is measured by the KLD between the distributions of the fake and true observations, *i.e.*, $D\left(f_z,f_y\right)$ (12). It is clear from Fig. 11 that the proposed detection scheme can detect completely stealthy, *i.e.*, $KLD=0$, measurement replacement-type attacks on the NCS. However, the detection delay increases with increasing stealthiness, *i.e.*, decreasing KLD.

## 6. Conclusion

We have studied the design of the quickest attack detection scheme by adding optimal random watermarking signals, where the attacker replaces the true observations by false data, and tries to cause damage to the NCS. There is a trade-off between the decrease in SADD and the increase in LQG control cost due to the addition of the watermarking signal. We have shown a strategy to find the optimum watermarking signal variance to minimize SADD for a given increase in LQG cost for a MISO system. We found that there is a single optimum eigenvalue and direction for the optimal watermarking signal variance. The relative magnitudes of the attack signal and the watermarking signal also play an important role in attack detection or attack stealthiness. The insights provided in the paper are useful to design a proper watermarking signal. The proposed sequential detection scheme can also be applied for replay attack detection after a few modifications. We have also compared the optimal CUSUM test with the optimal NP test to detect a deception attack and found the optimal CUSUM test to be quicker. In future, we will extend our sequential attack detection scheme to detect other kinds of attacks as well, such as additive attacks, DoS attacks, etc. We will also consider the real-time implementation of our attack detection scheme in a laboratory setup. Future work will also include the formulation of such attach detection problems as a dynamic two-player game between the control system designer and the attacker.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Appendix A. Proof of Theorem 1

Under the optimal CUSUM test, the likelihood ratio from (33) can be simplified using the chain rule of probability as

$$\frac{f_{\widetilde{\gamma}_k}\left(\bar{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}\right)f_{\mathbf{e}_{k-1}}(\mathbf{e}_{k-1})}{f_{\gamma_k}(\bar{\gamma}_k)f_{\mathbf{e}_{k-1}}(\mathbf{e}_{k-1})} \tag{A.1}$$

[$\mathbf{e}_k$ is iid and stationary, and $\gamma_k$ and $\mathbf{e}_{k-1}$ are uncorrelated]

$$= \frac{f_{\widetilde{\gamma}_k}\left(\bar{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}\right)}{f_{\gamma_k}(\bar{\gamma}_k)} \text{ [provided } f_{\mathbf{e}_{k-1}}(\mathbf{e}_{k-1}) \neq 0],$$

where $\bar{\gamma}_k = \gamma_k$ before the attack, and $\bar{\gamma}_k = \widetilde{\gamma}_k$ after the attack. The conditional mean $(\mu_{\widetilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}})$ and covariance $(\mathbf{\Sigma}_{\widetilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}})$ of $\widetilde{\gamma}_k$ are derived as follows.

The innovation signal under attack from (22) can be written as (A.2) after replacing $\mathbf{z}_k$ by (13),

$$\widetilde{\gamma}_k = \mathbf{w}_{a,k-1} + \mathbf{A}_a\mathbf{z}_{k-1} - \mathbf{C}(\mathbf{A}+\mathbf{BL})\hat{\mathbf{x}}^F_{k-1|k-1} - \mathbf{CBe}_{k-1}. \tag{A.2}$$

Applying (15), (20), (9) in (16) we can write,

$$\hat{\mathbf{x}}^F_{k|k} = (\mathbf{A}+\mathbf{BL})\hat{\mathbf{x}}^F_{k-1|k-1} + \mathbf{Be}_{k-1} + \mathbf{K}\widetilde{\gamma}_{k-1}. \tag{A.3}$$

Using (A.3) recursively we get,

$$\begin{aligned}\hat{\mathbf{x}}^F_{k|k} &= (\mathbf{A}+\mathbf{BL})^{k-1}\hat{\mathbf{x}}_{1|1} \\ &\quad + \sum_{i=1}^{k-1}(\mathbf{A}+\mathbf{BL})^{i-1}(\mathbf{Be}_{k-i-1}+\mathbf{K}\bar{\gamma}_{k-i}) \\ &\quad \text{where } \bar{\gamma}_k = \gamma_k \text{ for } k < \nu,\ \bar{\gamma}_k = \widetilde{\gamma}_k \text{ otherwise.}\end{aligned} \tag{A.4}$$

Applying (22) and (A.4) in (A.2) we get,

$$\begin{aligned}\widetilde{\gamma}_k &= \mathbf{w}_{a,k-1} + (\mathbf{A}_a\mathbf{C} - \mathbf{C}(\mathbf{A}+\mathbf{BL}))\Big((\mathbf{A}+\mathbf{BL})^{k-2}\hat{\mathbf{x}}_{1|1} \\ &\quad + \sum_{i=1}^{k-2}(\mathbf{A}+\mathbf{BL})^{i-1}\mathbf{Be}_{k-i-1} + \sum_{i=2}^{k-2}(\mathbf{A}+\mathbf{BL})^{i-1}\mathbf{K}\bar{\gamma}_{k-i}\Big) \\ &\quad - \mathbf{CBe}_{k-1} + (\mathbf{A}_a - \mathbf{C}(\mathbf{A}+\mathbf{BL})\mathbf{K})\bar{\gamma}_{k-1}.\end{aligned} \tag{A.5}$$

Since we have assumed that the system started at $k=-\infty$, and $(\mathbf{A}+\mathbf{BL})$ is strictly stable, we can say $(\mathbf{A}+\mathbf{BL})^{k-2} \approx \mathbf{0}$, and (A.5) will take the following form

$$\begin{aligned}\widetilde{\gamma}_k &= \mathbf{w}_{a,k-1} + (\mathbf{A}_a\mathbf{C} - \mathbf{C}(\mathbf{A}+\mathbf{BL})) \\ &\quad \left(\sum_{i=1}^{k-2}(\mathbf{A}+\mathbf{BL})^{i-1}\mathbf{Be}_{k-i-1} + \sum_{i=2}^{k-2}(\mathbf{A}+\mathbf{BL})^{i-1}\mathbf{K}\bar{\gamma}_{k-i}\right) \\ &\quad - \mathbf{CBe}_{k-1} + (\mathbf{A}_a - \mathbf{C}(\mathbf{A}+\mathbf{BL})\mathbf{K})\bar{\gamma}_{k-1}.\end{aligned} \tag{A.6}$$

Therefore,

$$\begin{aligned}\mu_{\widetilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}} &= E\left[\widetilde{\gamma}_k|\mathbf{z}_{k-1},\hat{\mathbf{x}}^F_{k-1|k-1},\mathbf{e}_{k-1}\right] \\ &= \mathbf{A}_a\mathbf{z}_{k-1} - \mathbf{C}(\mathbf{A}+\mathbf{BL})\hat{\mathbf{x}}^F_{k-1|k-1} - \mathbf{CBe}_{k-1}, \text{ and}\end{aligned} \tag{A.7}$$

$$\mathbf{\Sigma}_{\widetilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}} = cov\left(\widetilde{\gamma}_k|\mathbf{z}_{k-1},\hat{\mathbf{x}}^F_{k-1|k-1},\mathbf{e}_{k-1}\right) = \mathbf{Q}_a. \tag{A.8}$$

Furthermore, using (21) we obtain $E[\gamma_k] = 0$ and

$$\begin{aligned}&\gamma_k = \mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_{k|k-1} = \mathbf{C}\left(\mathbf{x}_k - \hat{\mathbf{x}}_{k|k-1}\right) + \mathbf{v}_k, \text{ and} \\ &\mathbf{\Sigma}_\gamma = E\left[\gamma_k\gamma_k^T\right] = \mathbf{CPC}^T + \mathbf{R}.\end{aligned} \tag{A.9}$$

## Appendix B. Proof of Corollary 1.1

The covariance matrix $(E[\widetilde{\gamma}_k\mathbf{e}_{k-1}^T])$ between $\widetilde{\gamma}_k$ (22) and $\mathbf{e}_{k-1}$ is evaluated as,

$$E\left[\widetilde{\gamma}_k\mathbf{e}_{k-1}^T\right] = E\left[-\mathbf{CBe}_{k-1}\mathbf{e}_{k-1}^T\right] = -\mathbf{CB\Sigma}_e, \tag{B.1}$$

since $\mathbf{e}_{k-1}$ is uncorrelated with $\mathbf{z}_k$ and $\hat{\mathbf{x}}^F_{k-1|k-1}$.

## Appendix C. Proof of Lemma 1

The variance of the innovation signal $(\mathbf{\Sigma}_{\widetilde{\gamma}})$ when the system is under attack is derived in this section. Using (22), 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 $\mathbf{\Sigma}_{\widetilde{\gamma}}$,

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

We first derive the expressions of $E\left[\hat{\mathbf{x}}^F_{k-1|k-1}\mathbf{z}_k^T\right]$ (C.7) and $E\left[\hat{\mathbf{x}}^F_{k-1|k-1}\left(\hat{\mathbf{x}}^F_{k-1|k-1}\right)^T\right]$ (C.10), and then use them to get the final expression of $\mathbf{\Sigma}_{\widetilde{\gamma}}$ (C.11). $E\left[\hat{\mathbf{x}}^F_{k-1|k-1}\mathbf{z}_k^T\right]$ is calculated using (15)–(17) and (20) as follows. First note that

$$\begin{aligned}&\hat{\mathbf{x}}^F_{k-1|k-1} = \mathbf{Kz}_{k-1} + A\hat{\mathbf{x}}^F_{k-2|k-2} + (\mathbf{I}_n - \mathbf{KC})\mathbf{Be}_{k-2}, \\ &\text{where } A = (\mathbf{I}_n - \mathbf{KC})(\mathbf{A}+\mathbf{BL}).\end{aligned} \tag{C.2}$$

$$\begin{aligned}&\text{We define } \mathbf{E}_{xz}(-k_0) \triangleq E\left[\hat{\mathbf{x}}^F_{k-k_0|k-k_0}\mathbf{z}_k^T\right], \\ &\qquad = E\Big[\Big(\mathbf{Kz}_{k-k_0} + A\hat{\mathbf{x}}^F_{k-k_0-1|k-k_0-1} \\ &\qquad\quad + (\mathbf{I}_n - \mathbf{KC})\mathbf{Be}_{k-k_0-1}\Big)\mathbf{z}_k^T\Big], \text{ [using (110)]} \\ &\qquad = \mathbf{KE}_{zz}(-k_0) + A\mathbf{E}_{xz}(-k_0-1),\end{aligned} \tag{C.3}$$

where $\mathbf{e}_{k-k_0-1}$ and $\mathbf{z}_k$ are uncorrelated, and $\mathbf{E}_{zz}(-k_0)$ is evaluated as follows.

$$\begin{aligned}&\mathbf{E}_{zz}(-k_0) = \mathbf{E}_{zz}(k_0) = E\left[\mathbf{z}_k\mathbf{z}_{k-k_0}^T\right], \\ &\mathbf{E}_{zz}(-1) = E\left[\mathbf{A}_a\mathbf{z}_{k-1}\mathbf{z}_{k-1}^T + \mathbf{w}_{a,k-1}\mathbf{z}_{k-1}^T\right] = \mathbf{A}_a\mathbf{E}_{zz}(0), \\ &\text{because } \mathbf{w}_{a,k} \text{ and } \mathbf{z}_k \text{ are uncorrelated. Similarly,} \\ &\mathbf{E}_{zz}(-2) = \mathbf{A}_a\mathbf{E}_{zz}(-1) = \mathbf{A}_a^2\mathbf{E}_{zz}(0), \text{ and} \\ &\mathbf{E}_{zz}(-k_0) = \mathbf{A}_a^{k_0}\mathbf{E}_{zz}(0).\end{aligned} \tag{C.4}$$

The system matrix $\mathbf{A}_a$ is assumed to be strictly 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 strictly stable $\mathbf{A}_a$,

$$\begin{aligned}&\mathbf{A}_a^{k_0} \to 0, \text{ as } k_0 \to \infty. \\ &\text{Therefore, } \mathbf{E}_{zz}(-k_0) \to 0, \text{ as } k_0 \to \infty.\end{aligned} \tag{C.5}$$

Using (C.3) and (C.4), we can write the expression of $\mathbf{E}_{xz}(-1)$ as

$$\begin{aligned}&\mathbf{E}_{xz}(-1) = \mathbf{KE}_{zz}(-1) + A\mathbf{E}_{xz}(-2) \\ &= \mathbf{KA}_a\mathbf{E}_{zz}(0) + A(\mathbf{KE}_{zz}(-2) + A\mathbf{E}_{xz}(-3)) \\ &\text{[after replacing } \mathbf{E}_{xz}(-2) \text{ using (111)]} \\ &= \mathbf{KA}_a\mathbf{E}_{zz}(0) + A\mathbf{KA}_a^2\mathbf{E}_{zz}(0) + A^2\mathbf{E}_{xz}(-3).\end{aligned} \tag{C.6}$$

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

$$\mathbf{E}_{xz}(-1) = \sum_{i=0}^{\infty} A^i\mathbf{KC}_a\mathbf{A}_a^{i+1}\mathbf{E}_{x_a}(0)\mathbf{C}_a^T. \tag{C.7}$$

$\mathbf{E}_{xz}(-1)$ can be evaluated numerically by taking a large number of terms for the summation (C.7), until the rest of the terms become negligible. $\mathbf{E}_{x^Fx^F}(0) = E\left[\hat{\mathbf{x}}^F_{k-1|k-1}\left(\hat{\mathbf{x}}^F_{k-1|k-1}\right)^T\right]$ is evaluated using (C.2) as

$$\begin{aligned}\mathbf{E}_{x^Fx^F}(0) &= \mathbf{K}E\left[\mathbf{z}_{k-1}\mathbf{z}_{k-1}^T\right]\mathbf{K}^T + AE\left[\hat{\mathbf{x}}^F_{k-2|k-2}\mathbf{z}^T_{k-1}\right]\mathbf{K}^T\\ &\quad + \left(AE\left[\hat{\mathbf{x}}^F_{k-2|k-2}\mathbf{z}^T_{k-1}\right]\mathbf{K}^T\right)^T\\ &\quad + AE\left[\hat{\mathbf{x}}^F_{k-2|k-2}\left(\hat{\mathbf{x}}^F_{k-2|k-2}\right)^T\right]A^T\\ &\quad + (\mathbf{I}_n - \mathbf{KC})\mathbf{B}E\left[\mathbf{e}_{k-2}\mathbf{e}^T_{k-2}\right]\mathbf{B}^T(\mathbf{I}_n - \mathbf{KC})^T. \end{aligned} \tag{C.8}$$

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

$$\begin{aligned}&A\mathbf{E}_{x^Fx^F}(0)A^T - \mathbf{E}_{x^Fx^F}(0) + \mathbf{K}\mathbf{E}_{zz}(0)\mathbf{K}^T\\ &\quad + A\mathbf{E}_{xz}(-1)\mathbf{K}^T + \left(A\mathbf{E}_{xz}(-1)\mathbf{K}^T\right)^T +\\ &(\mathbf{I}_n - \mathbf{KC})\mathbf{B}\boldsymbol{\Sigma}_e\mathbf{B}^T(\mathbf{I}_n - \mathbf{KC})^T = 0, \ [(112) \text{ used}]. \end{aligned} \tag{C.9}$$

$\mathbf{E}_{x^Fx^F}(0)$ is divided into two parts, $\boldsymbol{\Sigma}_{x^Fz}$ and $\boldsymbol{\Sigma}_{x^Fe}$ which are independent of the watermarking signal and the fake observations, respectively. $\boldsymbol{\Sigma}_{x^Fz}$ and $\boldsymbol{\Sigma}_{x^Fe}$ are the solution to the following Lyapunov equations,

$$\begin{aligned}&A\boldsymbol{\Sigma}_{x^Fz}A^T - \boldsymbol{\Sigma}_{x^Fz} + \mathbf{K}\mathbf{E}_{zz}(0)\mathbf{K}^T + A\mathbf{E}_{xz}(-1)\mathbf{K}^T\\ &\quad + \left(A\mathbf{E}_{xz}(-1)\mathbf{K}^T\right)^T = 0,\\ &A\boldsymbol{\Sigma}_{x^Fe}A^T - \boldsymbol{\Sigma}_{x^Fe} + (\mathbf{I}_n - \mathbf{KC})\mathbf{B}\boldsymbol{\Sigma}_e\mathbf{B}^T(\mathbf{I}_n - \mathbf{KC})^T = 0,\\ &\text{and } \mathbf{E}_{x^Fx^F}(0) = \boldsymbol{\Sigma}_{x^Fz} + \boldsymbol{\Sigma}_{x^Fe}. \end{aligned} \tag{C.10}$$

Using (C.4) and (C.10), we can rewrite the expression for $\boldsymbol{\Sigma}_{\tilde{\gamma}}$ as,

$$\begin{aligned}\boldsymbol{\Sigma}_{\tilde{\gamma}} &= \mathbf{E}_{zz}(0) - \mathbf{C}(\mathbf{A} + \mathbf{BL})\mathbf{E}_{xz}(-1)\\ &\quad - [\mathbf{C}(\mathbf{A} + \mathbf{BL})\mathbf{E}_{xz}(-1)]^T + \mathbf{CB}\boldsymbol{\Sigma}_e\mathbf{B}^T\mathbf{C}^T\\ &\quad + \mathbf{C}(\mathbf{A} + \mathbf{BL})\boldsymbol{\Sigma}_{x^Fz}(\mathbf{A} + \mathbf{BL})^T\mathbf{C}^T\\ &\quad + \mathbf{C}(\mathbf{A} + \mathbf{BL})\boldsymbol{\Sigma}_{x^Fe}(\mathbf{A} + \mathbf{BL})^T\mathbf{C}^T. \end{aligned} \tag{C.11}$$

## Appendix D. Proof of Corollary 1.2

We can simplify $\mathbf{E}_{xz}(-1)$ with the assumption that both $A$ and $\mathbf{A}_a$ are diagonalizable. If $A$ and $\mathbf{A}_a$ are diagonalizable, then the $i$th element of the expression for $\mathbf{E}_{xz}(-1)$, *i.e.*, $A^i\mathbf{K}\mathbf{A}_a^{i+1}\mathbf{E}_{zz}(0)$, will take the following form,

$$\begin{aligned}&\mathbf{U}_A\boldsymbol{\Sigma}^i_A\mathbf{U}_A^{-1}\mathbf{K}\mathbf{U}_a\boldsymbol{\Sigma}_a^i\mathbf{U}_a^{-1}\mathbf{A}_a\mathbf{E}_{zz}(0) \ [A \text{ and } \mathbf{A}_a \text{ replaced}\\ &\text{by eigenvalue decompositions, (49) and (50)]}\\ &= \mathbf{U}_A\boldsymbol{\Sigma}_A^i\mathbf{T}\boldsymbol{\Sigma}_a^i\mathbf{U}_a^{-1}\mathbf{A}_a\mathbf{E}_{zz}(0), \ [i = 0, \ \cdots \infty] \end{aligned} \tag{D.1}$$

where $\mathbf{T} = \mathbf{U}_A^{-1}\mathbf{K}\mathbf{U}_a$. $\mathbf{T}_a$ is defined as

$$\mathbf{T}_a \triangleq \sum_{i=0}^{\infty}\boldsymbol{\Sigma}_A^i\mathbf{T}\boldsymbol{\Sigma}_a^i. \tag{D.2}$$

The $jk$th element of the matrix $\mathbf{T}_a$ will be as follows

$$[\mathbf{T}_a]_{jk} = \sum_{i=0}^{\infty}[\mathbf{T}]_{jk}\lambda^i_{A,j}\lambda^i_{a,k} = \frac{[\mathbf{T}]_{jk}}{1 - \lambda_{A,j}\lambda_{a,k}} \tag{D.3}$$

where $[.]_{jk}$ denotes the $j$th row and $k$th column element of a matrix. $\lambda_{A,j}$ and $\lambda_{a,k}$ are the $j$th and $k$th diagonal element of the diagonal matrices $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_a$ respectively. We assume $A$ and $\mathbf{A}_a$ to be strictly stable, therefore, $|\lambda_{A,j}| < 1$ and $|\lambda_{a,k}| < 1$. $|.|$ denotes the absolute value of a scalar. Using (D.3), we can write

$$\mathbf{E}_{xz}(-1) = \mathbf{U}_A\mathbf{T}_a\mathbf{U}_a^{-1}\mathbf{A}_a\mathbf{E}_{zz}(0). \tag{D.4}$$

## Appendix E. Proof of Theorem 2

This section provides the proof of the Theorem 2 under the optimal CUSUM and sub-optimal CUSUM test. The KLDs for both the cases are derived using the general expression of KLD between two multivariate normal distributions given in [9]. Using (38), (39) and (A.2), and considering that $\mathbf{e}_k$ and $\mathbf{w}_{a,k}$ are uncorrelated with $\mathbf{z}_k$ and $\hat{\mathbf{x}}^F_{k|k}$, and also with each other, we can write,

$$\boldsymbol{\Sigma}_{\tilde{\gamma}} = \mathbf{Q}_a + E\left[\mu_{\tilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}}\mu^T_{\tilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}}\right]. \tag{E.1}$$

The expected KLD $E\left[D\left(f_{\tilde{\gamma}_k}, f_{\gamma_k}|\{\bar{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-1}\right)\right]$ under the optimal CUSUM test is derived as follows using [9], see (E.2).

$$\begin{aligned}&E\left[\tfrac{1}{2}\left(tr\left(\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\Sigma}_{\tilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}}\right) - m + \mu^T_{\tilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}}\boldsymbol{\Sigma}_\gamma^{-1}\mu_{\tilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}} - \log\frac{\left|\boldsymbol{\Sigma}_{\tilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}}\right|}{\left|\boldsymbol{\Sigma}_\gamma\right|}\right)\right]\\ &= \tfrac{1}{2}\left(-m + tr\left(\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\Sigma}_{\tilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}} + \boldsymbol{\Sigma}_\gamma^{-1}E\left[\mu_{\tilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}}\mu^T_{\tilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}}\right]\right) - \log\frac{\left|\boldsymbol{\Sigma}_{\tilde{\gamma}_k|\{\bar{\gamma}\}_1^{k-1},\{\mathbf{e}\}_1^{k-1}}\right|}{\left|\boldsymbol{\Sigma}_\gamma\right|}\right)\\ &= \tfrac{1}{2}\left\{tr\left(\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\Sigma}_{\tilde{\gamma}}\right) - m - \log\tfrac{|\mathbf{Q}_a|}{|\boldsymbol{\Sigma}_\gamma|}\right\}, [\text{using } (124)\&(39)]. \end{aligned} \tag{E.2}$$

Similarly, the KLD $D\left(f_{\tilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}}\right)$ under the sub-optimal CUSUM test will take the following form Duchi [9],

$$\frac{1}{2}\left(\log\frac{\left|\boldsymbol{\Sigma}_{\gamma_e}\right|}{\left|\boldsymbol{\Sigma}_{\tilde{\gamma}_e}\right|} - p - m + tr\left(\boldsymbol{\Sigma}_{\gamma_e}^{-1}\boldsymbol{\Sigma}_{\tilde{\gamma}_e}\right)\right). \tag{E.3}$$

The term $\log\frac{|\boldsymbol{\Sigma}_{\gamma_e}|}{|\boldsymbol{\Sigma}_{\tilde{\gamma}_e}|}$ is evaluated as follows,

$$\left|\boldsymbol{\Sigma}_{\gamma_e}\right| = \left|\boldsymbol{\Sigma}_e\right|\left|\boldsymbol{\Sigma}_\gamma\right|, \ [\text{using } (42)] \tag{E.4}$$

$$\left|\boldsymbol{\Sigma}_{\tilde{\gamma}_e}\right| = \left|\boldsymbol{\Sigma}_e\right|\left|\boldsymbol{\Sigma}_{\tilde{\gamma}} - \mathbf{CB}\boldsymbol{\Sigma}_e\mathbf{B}^T\mathbf{C}^T\right|, \ [\text{using } (43)]. \tag{E.5}$$

$$\text{Therefore, } \log\frac{\left|\boldsymbol{\Sigma}_{\gamma_e}\right|}{\left|\boldsymbol{\Sigma}_{\tilde{\gamma}_e}\right|} = -\log\frac{\left|\boldsymbol{\Sigma}_{\tilde{\gamma}} - \mathbf{CB}\boldsymbol{\Sigma}_e\mathbf{B}^T\mathbf{C}^T\right|}{\left|\boldsymbol{\Sigma}_\gamma\right|}. \tag{E.6}$$

The term $tr\left(\boldsymbol{\Sigma}_{\gamma_e}^{-1}\boldsymbol{\Sigma}_{\tilde{\gamma}_e}\right)$ is evaluated using (42) and (43) as,

$$tr\left(\boldsymbol{\Sigma}_{\gamma_e}^{-1}\boldsymbol{\Sigma}_{\tilde{\gamma}_e}\right) = tr\left(\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\Sigma}_{\tilde{\gamma}} + \boldsymbol{\Sigma}_e^{-1}\boldsymbol{\Sigma}_e\right) = tr\left(\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\Sigma}_{\tilde{\gamma}}\right) + p \tag{E.7}$$

Applying (E.6) and (E.7) in (E.3), we get the final expression of the KLD $D\left(f_{\tilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}}\right)$ under the sup-optimal CUSUM test as

$$\frac{1}{2}\left\{tr\left(\boldsymbol{\Sigma}_\gamma^{-1}\boldsymbol{\Sigma}_{\tilde{\gamma}}\right) - m - \log\frac{\mid\boldsymbol{\Sigma}_{\tilde{\gamma}} - \mathbf{CB}\boldsymbol{\Sigma}_e\mathbf{B}^T\mathbf{C}^T\mid}{\mid\boldsymbol{\Sigma}_\gamma\mid}\right\}. \tag{E.8}$$

## Appendix F. Proof of Lemma 2

This section provides the derivation of the expression of $\sigma^2_{\tilde{\gamma}}$ for the MISO system. The model parameters of the fake measurement generation system (13)) for the MISO system will be as follows.

$$\mathbf{A}_a = \rho, \ \mathbf{Q}_a = \left(1 - \rho^2\right)\sigma_z^2, \text{and } \mathbf{E}_{zz}(0) = \sigma_z^2. \tag{F.1}$$

To evaluate $\sigma_{\tilde{\gamma}}^2$, we derive the expression for $\mathbf{E}_{xz}(-1)$ for a MISO system using (45) as

$$\begin{aligned}\mathbf{E}_{xz}(-1) &= \sum_{i=0}^{\infty} A^i \mathbf{K}\mathbf{A}_a^{i+1}\mathbf{E}_{zz}(0)\\ &= \sum_{i=0}^{\infty} A^i \mathbf{K}\rho^{i+1}\sigma_z^2,\ [\mathbf{E}_{zz}(0) = \sigma_z^2,\ \ \mathbf{A}_a = \rho]\\ &= [\mathbf{I}_n - \rho A]^{-1}\mathbf{K}\rho\sigma_z^2,\ [A \text{ is strictly stable, } \rho < 1]. \end{aligned} \tag{F.2}$$

$\sigma_{\tilde{\gamma}}^2$ will be as follows,

$$\begin{aligned}\sigma_{\tilde{\gamma}}^2 &= \sigma_z^2 - 2\mathbf{C}(\mathbf{A}+\mathbf{BL})\mathbf{E}_{xz}(-1) + \mathbf{CB}\boldsymbol{\Sigma}_e\mathbf{B}^T\mathbf{C}^T\\ &\quad + \mathbf{C}(\mathbf{A}+\mathbf{BL})\boldsymbol{\Sigma}_{x^F z}(\mathbf{A}+\mathbf{BL})^T\mathbf{C}^T\\ &\quad + \mathbf{C}(\mathbf{A}+\mathbf{BL})\boldsymbol{\Sigma}_{x^F e}(\mathbf{A}+\mathbf{BL})^T\mathbf{C}^T \ [\text{using (44)}],\end{aligned} \tag{F.3}$$

where $\boldsymbol{\Sigma}_{x^F z}$ and $\boldsymbol{\Sigma}_{x^F e}$ are derived from (46) and (47) respectively as follows.

$$\boldsymbol{\Sigma}_{x^F z} = \boldsymbol{\Sigma}_{x^F}^z\sigma_z^2 \tag{F.4}$$

where $\boldsymbol{\Sigma}_{x^F}^z$ is the solution to the following Lyapunov equation,

$$\begin{aligned}&A\boldsymbol{\Sigma}_{x^F}^z A^T - \boldsymbol{\Sigma}_{x^F}^z + \mathbf{K}\mathbf{K}^T + A[\mathbf{I}_n - \rho A]^{-1}\mathbf{K}\mathbf{K}^T\rho\\ &\quad + \left[A[\mathbf{I}_n - \rho A]^{-1}\mathbf{K}\mathbf{K}^T\rho\right]^T = 0.\end{aligned} \tag{F.5}$$

$\boldsymbol{\Sigma}_{x^F e}$ is the solution to the following Lyapunov equation,

$$A\boldsymbol{\Sigma}_{x^F e}A^T - \boldsymbol{\Sigma}_{x^F e} + (\mathbf{I}_n - \mathbf{KC})\mathbf{B}\boldsymbol{\Sigma}_e\mathbf{B}^{\mathbf{T}}(\mathbf{I}_n - \mathbf{KC})^T = 0. \tag{F.6}$$

Using (F.2) and (F.4), the expression for $\sigma_{\tilde{\gamma}}^2$ (F.3) can be rearranged as follows.

$$\begin{aligned}\sigma_{\tilde{\gamma}}^2 &= \Big(1 - 2\mathbf{C}(\mathbf{A}+\mathbf{BL})(\mathbf{I}_n - \rho A)^{-1}\mathbf{K}\rho\\ &+\mathbf{C}(\mathbf{A}+\mathbf{BL})\boldsymbol{\Sigma}_{x^F}^z(\mathbf{A}+\mathbf{BL})^T\mathbf{C}^T\Big)\sigma_z^2\\ &\quad + \Big(\mathbf{C}(\mathbf{A}+\mathbf{BL})\boldsymbol{\Sigma}_{xe}(\mathbf{A}+\mathbf{BL})^T\mathbf{C}^T + \mathbf{CB}\boldsymbol{\Sigma}_e\mathbf{B}^T\mathbf{C}^T\Big)\\ &= M_z\sigma_z^2 + M_t\end{aligned} \tag{F.7}$$

The scalar quantity $M_t$ can be rearranged as follows.

$$\begin{aligned}M_t &= \Bigg(\sum_{t=0}^{\infty}\mathbf{C}(\mathbf{A}+\mathbf{BL})A^t(\mathbf{I}_n - \mathbf{KC})\mathbf{B}\boldsymbol{\Sigma}_e\mathbf{B}^{\mathbf{T}}(\mathbf{I}_n - \mathbf{KC})^T\\ &\qquad \times \left[A^T\right]^t(\mathbf{A}+\mathbf{BL})^T\mathbf{C}^T\Bigg) + \mathbf{CB}\boldsymbol{\Sigma}_e\mathbf{B}^T\mathbf{C}^T\\ &= \mathrm{tr}\Bigg(\sum_{t=0}^{\infty}\mathbf{B}^{\mathbf{T}}(\mathbf{I}_n - \mathbf{KC})^T\left[A^T\right]^t(\mathbf{A}+\mathbf{BL})^T\mathbf{C}^T\mathbf{C}(\mathbf{A}+\mathbf{BL})\\ &A^t(\mathbf{I}_n - \mathbf{KC})\mathbf{B}\boldsymbol{\Sigma}_e + \mathbf{B}^T\mathbf{C}^T\mathbf{CB}\boldsymbol{\Sigma}_e\Bigg) = \mathrm{tr}(M_e\boldsymbol{\Sigma}_e),\end{aligned} \tag{F.8}$$

$$\text{where } M_e = \mathbf{B}^{\mathbf{T}}(\mathbf{I}_n - \mathbf{KC})^T\boldsymbol{\Sigma}_{x^F}^e(\mathbf{I}_n - \mathbf{KC})\mathbf{B} + \mathbf{B}^T\mathbf{C}^T\mathbf{CB}. \tag{F.9}$$

$\boldsymbol{\Sigma}_{x^F}^e$ is the solution to the following Lyapunov equation,

$$A^T\boldsymbol{\Sigma}_{x^F}^e A - \boldsymbol{\Sigma}_{x^F}^e + (\mathbf{A}+\mathbf{BL})^T\mathbf{C}^T\mathbf{C}(\mathbf{A}+\mathbf{BL}) = 0. \tag{F.10}$$

Finally, we can write $\sigma_{\tilde{\gamma}}^2$ as

$$\sigma_{\tilde{\gamma}}^2 = M_z\sigma_z^2 + \mathrm{tr}(M_e\boldsymbol{\Sigma}_e). \tag{F.11}$$

## Appendix G. Proof of Theorem 4

The covariance matrix of the watermarking signal is decomposed using eigenvalue decomposition as follows,

$$\boldsymbol{\Sigma}_e = \mathbf{V}_e\boldsymbol{\Lambda}_e\mathbf{V}_e^T \tag{G.1}$$

where $\mathbf{V}_e$ and $\boldsymbol{\Lambda}_e$ are the eigenvector matrix and the diagonal eigenvalue matrix. In this section, we will prove that KLD is convex with respect to the elements of $\boldsymbol{\Lambda}_e$ for a fixed $\mathbf{V}_e$. We formulate the optimization problem as follows.

$$\begin{aligned}&\max_{\boldsymbol{\Lambda}_e} f(\boldsymbol{\Lambda}_e) = \ E\Big[D\Big(f_{\tilde{\gamma}_k}, f_{\gamma_k}|\{\tilde{\gamma}\}_1^{k-1}, \{\mathbf{e}\}_1^{k-1}\Big)\Big]\text{or}\\ &\max_{\boldsymbol{\Lambda}_e} f(\boldsymbol{\Lambda}_e) = \ D\Big(f_{\tilde{\gamma}_k,\mathbf{e}_{k-1}}, f_{\gamma_k,\mathbf{e}_{k-1}}\Big)\end{aligned} \tag{G.2}$$

$$\text{s.t. } \Delta LQG \le J \tag{G.3}$$

$$\text{and } \lambda_{e,i} \ge 0, \forall i. \tag{G.4}$$

The proof for the optimal CUSUM case is as follows.

Observing (53) and (44), we can say that maximizing the expected KLD with respect to $\boldsymbol{\Sigma}_e$ is the same as maximizing the following portion of the expected KLD expression which is only dependent on $\boldsymbol{\Sigma}_e$.

$$f(\boldsymbol{\Sigma}_e) = \mathbf{C}(\mathbf{A}+\mathbf{BL})\boldsymbol{\Sigma}_{x^F e}(\mathbf{A}+\mathbf{BL})^T\mathbf{C}^T + \mathbf{CB}\boldsymbol{\Sigma}_e\mathbf{B}^T\mathbf{C}^T \tag{G.5}$$

where $\boldsymbol{\Sigma}_{x^F e}$ is given by (47). Putting the solution of (47) in (G.5), we get,

$$\begin{aligned}f(\boldsymbol{\Sigma}_e) &= \mathbf{C}(\mathbf{A}+\mathbf{BL})\Bigg(\sum_{t=0}^{\infty}A^t(\mathbf{I}_n - \mathbf{KC})\mathbf{B}\boldsymbol{\Sigma}_e\mathbf{B}^T\\ &(\mathbf{I}_n - \mathbf{KC})^T\left[A^T\right]^t\Bigg)(\mathbf{A}+\mathbf{BL})^T + \mathbf{CB}\boldsymbol{\Sigma}_e\mathbf{B}^T\mathbf{C}^T\\ &= tr\Big(\Big(\mathbf{B}^T(\mathbf{I}_n - \mathbf{KC})^T L_e(\mathbf{I}_n - \mathbf{KC})\mathbf{B} + \mathbf{B}^T\mathbf{C}^T\mathbf{CB}\Big)\boldsymbol{\Sigma}_e\Big)\\ &= tr(\mathbf{H}_{KLD}\boldsymbol{\Sigma}_e),\end{aligned} \tag{G.6}$$

where $L_e$ is the solution to the following Lyapunov equation

$$A^T L_e A - L_e + (\mathbf{A}+\mathbf{BL})^T\mathbf{C}^T\mathbf{C}(\mathbf{A}+\mathbf{BL}) = 0, \text{and} \tag{G.7}$$

$$\mathbf{H}_{KLD} = \mathbf{B}^T(\mathbf{I}_n - \mathbf{KC})^T L_e(\mathbf{I}_n - \mathbf{KC})\mathbf{B} + \mathbf{B}^T\mathbf{C}^T\mathbf{CB}. \tag{G.8}$$

Using (G.6) and (G.1), we can rewrite the cost function as follows

$$f(\boldsymbol{\Lambda}_e) = tr\Big(\mathbf{V}_e^T\mathbf{H}_{KLD}\mathbf{V}_e\boldsymbol{\Lambda}_e\Big) \tag{G.9}$$

which represents a line in the $p$ dimensional hyperplane. Therefore, the cost function is convex in nature.

The proof for the sub-optimal CUSUM case is as follows. We have replaced all the $\mathbf{B}$ matrices by $\mathbf{B}_e$ where $\mathbf{B}_e = \mathbf{B}\mathbf{V}_e$ and $\boldsymbol{\Sigma}_e$ by $\boldsymbol{\Lambda}_e$ to keep the structure of the KLD and $\sigma_{\tilde{\gamma}}^2$ expressions as (64) and (66) respectively.

$$\begin{aligned}f(\boldsymbol{\Lambda}_e) &= \frac{1}{2}\left(\frac{M_z\sigma_z^2 + \sum_{i=1}^{n}[\mathbf{M}_{e\lambda}]_{ii}\lambda_{e,i}}{\sigma_{\gamma}^2}\right)\\ &\quad - \frac{1}{2}\log\left(\frac{M_z\sigma_z^2 + \sum_{i=1}^{n}[\mathbf{M}_{em}]_{ii}\lambda_{e,i}}{\sigma_{\gamma}^2}\right)\end{aligned} \tag{G.10}$$

$$\begin{aligned}&\text{where } \mathbf{M}_{em} = \mathbf{B}_e^T(\mathbf{I}_n - \mathbf{KC})^T\boldsymbol{\Sigma}_{x^F}^e(\mathbf{I}_n - \mathbf{KC})\mathbf{B}_e, \text{ and}\\ &\mathbf{M}_{e\lambda} = \mathbf{B}_e^T(\mathbf{I}_n - \mathbf{KC})^T\boldsymbol{\Sigma}_{x^F}^e(\mathbf{I}_n - \mathbf{KC})\mathbf{B}_e + \mathbf{B}_e^T\mathbf{C}^T\mathbf{CB}_e.\end{aligned} \tag{G.11}$$

The $\boldsymbol{\Sigma}_{x^F}^e$ is the same as in (F.10). The first derivative of the cost function with respect to the $j$th eigenvalue $\lambda_{e,j}$ is as follows,

$$\frac{\partial}{\partial\lambda_{e,j}}f(\boldsymbol{\Lambda}_e) = \frac{1}{2\sigma_{\gamma}^2}[\mathbf{M}_{e\lambda}]_{jj} \tag{G.12}$$

$$- \frac{1}{2}\frac{1}{M_z\sigma_z^2 + \sum_{i=1}^{n}[\mathbf{M}_{em}]_{ii}\lambda_{e,i}}[\mathbf{M}_{em}]_{jj}. \tag{G.12}$$

The second derivative of the cost function is as follows,

$$\frac{\partial}{\partial \lambda_{e,i}} \frac{\partial}{\partial \lambda_{e,j}} f(\mathbf{\Lambda}_e) = \frac{1}{2}[\mathbf{M}_{em}]_{ii}[\mathbf{M}_{em}]_{jj} t_f^2, \text{ and}$$
$$t_f = \frac{1}{M_z \sigma_z^2 + \sum_{i=1}^{n} [\mathbf{M}_{em}]_{ii} \lambda_{e,i}} \tag{G.13}$$

where $\frac{\partial}{\partial \lambda_{e,i}} \frac{\partial}{\partial \lambda_{e,j}} f(\mathbf{\Lambda}_e)$ is the $ij$th element of the Hessian matrix $\mathbf{H}_s = \nabla^2_{\mathbf{\Lambda}_e} f(\mathbf{\Lambda}_e)$. From (G.13), it is clear that each column of $\mathbf{H}_s$ is linearly dependent on any other column of the matrix. This means that we have all eigenvalues except one to be zero. Therefore, determinants of all the principle minors of $\mathbf{H}_s$ are zero. Also, the diagonal elements of $\mathbf{H}_s$ are non-zero. So, we can conclude that KLD is convex in $\mathbf{\Lambda}_e$.

Since the cost function under both the tests are convex, the optimum $\mathbf{\Lambda}_e$, which maximizes the expected KLD or the KLD, will be on one of the vertices of the feasible region provided by (G.3) and (G.4). That is possible when the optimum $\mathbf{\Lambda}_e$ contains only one non-zero element. This property of the convex function over a polyhedron set can be proved using Jensen's inequality.

## Appendix H. System parameters

For both the systems, $ARL_h = 1000$.

**System-A parameters**:

$$\mathbf{A} = \begin{bmatrix} 0.75 & 0.2 \\ 0.2 & 1.0 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 0.9 & 0.5 \\ 0.1 & 1.2 \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} 1.0 & -1.0 \end{bmatrix}$$
$$\mathbf{Q} = diag\begin{bmatrix} 1 & 1 \end{bmatrix} \qquad \mathbf{R} = 1 \qquad \mathbf{W} = diag\begin{bmatrix} 1 & 2 \end{bmatrix}$$
$$\mathbf{U} = diag\begin{bmatrix} 0.4 & 0.7 \end{bmatrix} \qquad \sigma_z^2 = 10 \qquad \rho = 0.5$$

**System-B parameters**:

$$\mathbf{A} = \begin{bmatrix} 0.968 & 0 & 0.082 & 0 \\ 0 & 0.978 & 0 & 0.064 \\ 0 & 0 & 0.917 & 0 \\ 0 & 0 & 0 & 0.935 \end{bmatrix} \mathbf{B} = \begin{bmatrix} 0.164 & 0.004 \\ 0.002 & 0.124 \\ 0 & 0.092 \\ 0.060 & 0 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \end{bmatrix} \qquad \mathbf{R} = diag\begin{bmatrix} 0.5 & 0.5 \end{bmatrix}$$
$$\mathbf{Q} = diag\begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix} \qquad \mathbf{U} = diag\begin{bmatrix} 2 & 2 \end{bmatrix}$$
$$\mathbf{W} = diag\begin{bmatrix} 5 & 5 & 1 & 1 \end{bmatrix} \qquad \mathbf{Q}_a = diag\begin{bmatrix} 5 & 5 \end{bmatrix}$$
$$\mathbf{A}_a = diag\begin{bmatrix} 0.4 & 0.2 & 0.2 & 0.7 \end{bmatrix}$$

## References

[1] M. Abrams, J. Weiss, Malicious Control System Cyber Security Attack Case Study Maroochy Water Services, Australia, MITRE Corp USA 253 (August) (2008) 73–82.

[2] P.T. Boggs, J.W. Tolle, Sequential Quadratic Programming, Acta Numer. 4 (1995) (1995) 1–51, doi:10.1017/S0962492900002518.

[3] S. Boyd, L. Vandenberghe, Convex optimization, Cambridge university press, 2004.

[4] M.M. Bruce, Estimation of variance by a recursive equation 5465 (1969).

[5] A. Cardenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, S. Sastry, Challenges for Securing Cyber Physical Systems, 2009, 10.1016/0960-2593(92)90002-5

[6] A.A. Cárdenas, S. Amin, S. Sastry, Secure control: Towards survivable cyber-physical systems, Proc. - Int. Conf. Distrib. Comput. Syst. (2008) 495–500, doi:10.1109/ICDCS.Workshops.2008.40.

[7] Y. Chen, S. Kar, J.M. Moura, Cyber-Physical Attacks with Control Objectives, IEEE Trans. Automat. Contr. 63 (5) (2018) 1418–1425, doi:10.1109/TAC.2017.2741778.

[8] D. Du, X. Li, W. Li, R. Chen, M. Fei, L. Wu, ADMM-Based Distributed State Estimation of Smart Grid under Data Deception and Denial of Service Attacks, IEEE Trans. Syst. Man, Cybern. Syst. 49 (8) (2019) 1698–1711, doi:10.1109/TSMC.2019.2896292.

[9] J. Duchi, Derivations for Linear Algebra and Optimization, Berkeley, Calif. (2007) 1–13.

[10] C. Fang, Y. Qi, P. Cheng, W.X. Zheng, Optimal periodic watermarking schedule for replay attack detection in cyberphysical systems, Automatica 112 (2020) 108698, doi:10.1016/j.automatica.2019.108698.

[11] H. Fawzi, P. Tabuada, S. Diggavi, Secure estimation and control for cyber-physical systems under adversarial attacks, IEEE Trans. Automat. Contr. 59 (6) (2014) 1454–1467, doi:10.1109/TAC.2014.2303233.

[12] A. Forsgren, P.E. Gill, M.H. Wright, Interior methods for nonlinear optimization, SIAM Rev. 44 (4) (2002) 525–597, doi:10.1137/S0036144502414942.

[13] N. Forti, G. Battistelli, L. Chisci, S. Li, B. Wang, B. Sinopoli, Distributed Joint Attack Detection and Secure State Estimation, IEEE Trans. Signal Inf. Process. over Networks 4 (1) (2018) 96–110, doi:10.1201/9781420007275-4.

[14] X. Ge, Q.L. Han, M. Zhong, X.M. Zhang, Distributed Krein space-based attack detection over sensor networks under deception attacks, Automatica 109 (2019) 108557, doi:10.1016/j.automatica.2019.108557.

[15] M. Ghaderi, K. Gheitasi, W. Lucia, A blended active detection strategy for false data injection attacks in cyber-physical systems, IEEE Transactions on Control of Network Systems 8 (1) (2020) 168–176.

[16] J. Giraldo, 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, doi:10.1145/3314058.3318166.

[17] V. Girardin, V. Konev, S. Pergamenchtchikov, Kullback-Leibler Approach to CUSUM Quickest Detection Rule for Markovian Time Series, Seq. Anal. 37 (3) (2018) 322–341, doi:10.1080/07474946.2018.1548846.

[18] K.H. Johansson, J.L.R. Nunes, The Quadruple-Tank Process: A Multivariable Laboratory Process with an Adjustable Zero, Proc. Am. Control Conf. 8 (3) (2000) 456–465, doi:10.1109/ACC.1998.702986.

[19] W.H. Ko, B. Satchidanandan, P.R. Kumar, Dynamic watermarking-based defense of transportation cyber-physical systems, ACM Trans. Cyber-Physical Syst. 4 (1) (2019), doi:10.1145/3361700.

[20] J. Kuhn, M. Mandjes, T. Taimre, Practical aspects of false alarm control for change point detection: Beyond average run length, Methodology and Computing in Applied Probability 21 (1) (2019) 25–42.

[21] T.L. Lai, Information bounds and quick detection of parameter changes in stochastic systems, IEEE Trans. Inf. Theory 44 (7) (1998) 2917–2929, doi:10.1109/18.737522.

[22] R. Langner, Stuxnet: Dissecting a cyberwarfare weapon, IEEE Secur. Priv. 9 (3) (2011) 49–51, doi:10.1109/MSP.2011.67.

[23] P. Li, D. Ye, Measurement-based optimal stealthy attacks on remote state estimation, IEEE Transactions on Information Forensics and Security (2022).

[24] X.-J. Li, X.-Y. Shen, A data-driven attack detection approach for dc servo motor systems based on mixed optimization strategy, IEEE Transactions on Industrial Informatics 16 (9) (2020) 5806–5813, doi:10.1109/TII.2019.2960616.

[25] Y.-C. Liu, G. Bianchin, F. Pasqualetti, Secure trajectory planning against undetectable spoofing attacks, Automatica 112 (2020) 108655.

[26] Y. Mo, R. Chabukswar, B. Sinopoli, Detecting integrity attacks on SCADA systems, IEEE Trans. Control Syst. Technol. 22 (4) (2014) 1396–1407, doi:10.1109/TCST.2013.2280899.

[27] Y. Mo, B. Sinopoli, Secure control against replay attacks, 2009 47th Annu. Allert. Conf. Commun. Control. Comput. Allert. 2009 (2009) 911–918, doi:10.1109/ALLERTON.2009.5394956.

[28] Y. Mo, S. Weerakkody, B. Sinopoli, Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs, IEEE Control Syst. 35 (1) (2015) 93–109, doi:10.1109/MCS.2014.2364724.

[29] A. Mousavi, K. Aryankia, R.R. Selmic, A distributed fdi cyber-attack detection in discrete-time nonlinear multi-agent systems using neural networks, European Journal of Control 66 (2022) 100646.

[30] E. Mousavinejad, F. Yang, Q.L. Han, L. Vlacic, A novel cyber attack detection method in networked control systems, IEEE Trans. Cybern. 48 (11) (2018) 3254–3264, doi:10.1109/TCYB.2018.2843358.

[31] C. Murguia, J. Ruths, CUSUM and chi-squared attack detection of compromised sensors, in: IEEE Conf. Control Appl., 2016, pp. 474–480, doi:10.1109/cca.2016.7587875.

[32] A. Naha, A. Teixeira, A. Ahlen, S. Dey, Sequential detection of replay attacks, IEEE Transactions on Automatic Control (Early Access) (2022), doi:10.1109/TAC.2022.3174004.

[33] G. Park, C. Lee, H. Shim, Y. Eun, K.H. Johansson, Stealthy Adversaries against Uncertain Cyber-Physical Systems: Threat of Robust Zero-Dynamics Attack, IEEE Trans. Automat. Contr. 64 (12) (2019) 4907–4919, doi:10.1109/TAC.2019.2903429.

[34] F. Pasqualetti, F. Dorfler, F. Bullo, Attack detection and identification in cyber-physical systems, IEEE Trans. Automat. Contr. 58 (11) (2013) 2715–2729, doi:10.1109/TAC.2013.2266831.

[35] F. Pasqualetti, F. Dorfler, F. Bullo, Control-theoretic methods for cyberphysical security: Geometric principles for optimal cross-layer resilient control systems, IEEE Control Systems Magazine 35 (1) (2015) 110–127.

[36] B. Ramasubramanian, M. Rajan, M.G. Chandra, R. Cleaveland, S.I. Marcus, Resilience to denial-of-service and integrity attacks: A structured systems approach, European Journal of Control 63 (2022) 61–69.

[37] S. Salimi, S. Dey, A. Ahlen, Sequential detection of deception attacks in networked control systems with watermarking, 2019 18th Eur. Control Conf. ECC 2019 (2019) 883–890, doi:10.23919/ECC.2019.8796303.

[38] B. Satchidanandan, P.R. Kumar, Dynamic Watermarking: Active Defense of Networked CyberPhysical Systems, Proc. IEEE 105 (2) (2017) 219–240, doi:10.1109/JPROC.2016.2575064.

[39] B. Satchidanandan, P.R. Kumar, On the Design of Security-Guaranteeing Dynamic Watermarks, IEEE Control Syst. Lett. 4 (2) (2020) 307–312, doi:10.1109/LCSYS.2019.2925278.

[40] A.N. Shiryaev, On optimum methods in quickest detection problems, Theory Probab. Its Appl. 8 (1) (1963) 22–46.

[41] J.J. Stapleton, Security without obscurity: A guide to confidentiality, authentication, and integrity, cRc Press, 2014.

[42] A. Tartakovsky, I. Nikiforov, M. Basseville, Sequential analysis: Hypothesis testing and changepoint detection, 2014, doi:10.1080/02664763.2015.1015813.

[43] A.G. Tartakovsky, On Asymptotic Optimality in Sequential Changepoint Detection: Non-iid Case, IEEE Trans. Inf. Theory 63 (6) (2017) 3433–3450, doi:10.1109/TIT.2017.2683496.

[44] A.G. Tartakovsky, V.V. Veeravalli, Asymptotically optimal quickest change detection in distributed sensor systems, Seq. Anal. 27 (4) (2008) 441–475, doi:10.1080/07474940802446236.

[45] R. Tunga, C. Murguia, J. Ruths, Tuning Windowed Chi-Squared Detectors for Sensor Attacks, in: Proc. Am. Control Conf., 2018, pp. 1752–1757, doi:10.23919/ACC.2018.8431073.

[46] D.I. Urbina, J. Giraldo, A.A. Cardenas, N.O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, H. Sandberg, Limiting the impact of stealthy attacks on Industrial Control Systems, in: Proc. ACM Conf. Comput. Commun. Secur., 2016, pp. 1092–1105, doi:10.1145/2976749.2978388.

[47] S. Weerakkody, O. Ozel, B. Sinopoli, A Bernoulli-Gaussian physical watermark for detecting integrity attacks in control systems, 55th Annu. Allert. Conf. Commun. Control. Comput. Allert. 2017 2018-Janua (Iid) (2018) 966–973, doi:10.1109/ALLERTON.2017.8262842.

[48] B. Yaghooti, R. Romagnoli, B. Sinopoli, Physical watermarking for replay attack detection in continuous-time systems, European Journal of Control 62 (2021) 57–62.

[49] G.-Y. Yang, X.-J. Li, Complete stealthiness false data injection attacks against dynamic state estimation in cyber-physical systems, Information Sciences 586 (2022) 408–423.

[50] M.H. Yılmaz, H. Arslan, A survey: Spoofing attacks in physical layer security, in: 2015 IEEE 40th Local Computer Networks Conference Workshops (LCN Workshops), IEEE, 2015, pp. 812–817.

[51] J. Zhang, R.S. Blum, L.M. Kaplan, X. Lu, Functional forms of optimum spoofing attacks for vector parameter estimation in quantized sensor networks, IEEE Transactions on Signal Processing 65 (3) (2016) 705–720.

[52] T.-Y. Zhang, D. Ye, False data injection attacks with complete stealthiness in cyber–physical systems: A self-generated approach, Automatica 120 (2020) 109117.