Quickest Detection of Deception Attacks in Networked Control Systems with Physical Watermarking

Arunava Naha¹, André Teixeira¹, Anders Ahlén¹ and Subhrakanti Dey²

Abstract—In this paper, we propose and analyze an attack detection scheme for securing the physical layer of a networked control system 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) 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.

Index Terms—CUSUM test, cyber-physical system, deception attack, Kullback–Leibler divergence, linear quadratic Gaussian control, networked control system, physical watermarking, resilient attack detection

I. INTRODUCTION

Large distributed networked control systems (NCS) are getting deployed in various sectors such as manufacturing units, transportation systems, power systems, robotics, etc. [1]. Such cyber-physical systems (CPS) consist of embedded software, processors and other physical components. The components of CPS may be distributed over a large area, and communicate with each other via wired or wireless links. 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, such as, e.g., the Stuxnet attack [2], the attack on the sewage systems in Australia [3], the attack on the Davis-Besse nuclear power plant in Ohio, USA [4]. Attacks on such systems can cause loss of production, financial loss, a threat to human safety, etc. Securing CPS is a great challenge. The cyber layer is usually secured by employing cryptography, digital watermarking, etc. However, these measures cannot ensure the safety of the physical layer of the system.

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 [5]. In the deception attack, the adversary feeds the NCS with false data either by replacing or distorting the true observations and/or the control inputs [1], [5]. The attacker always tries to statistically match the fake data to the real ones to remain stealthy. 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 at some later point in time. Such an attack strategy is called a replay attack [5]. In the DoS attack, the attacker makes the data unavailable maybe by jamming the wireless network [6]. In both the attack strategies, the attacker’s objective is to make the system unstable or force the system to operate at a state outside it’s desired normal behaviour, and at the same time to remain stealthy as long as possible to cause maximum damage [1], [5], [6]. In this paper, we have studied mainly a specific scenario of deception attacks, where the attacker hijacks the sensor nodes and feeds random but stationary fake observations to the state estimator. The noise and the uncertainty in the system always facilitate the attacker to remain stealthy. We also assume that the attacker has complete knowledge about the system, and controller parameters and knows the statistical properties of the noise and observations.

A. Related Work

Several different approaches are found in the literature to secure CPS from the 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 [7]–[9]. In [10], [11], the authors have studied different attack strategies which will be useful to design more resilient defence strategies. The defence strategies employed for attack detections 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 \([12]–[14]\). For example, a set membership filter-based algorithm is used in \([13]\) to detect malicious data injection attacks in the NCS, a two-stage distributed deception attack detection mechanism is published in \([14]\) based on the residual analysis of the Krein state-space model and locally distributed estimators. The passive detection schemes, in general, have an unsatisfactory probability of detection in the presence of noise and uncertainties.

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 \([1]–[5], [15]–[18]\). In our paper, we follow this approach to design a resilient deception attack detection scheme. The idea of physical watermarking is analogous to the digital watermarking, which is used to authenticate the actual owner of a digital content. In \([15]\), 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 to the linear quadratic Gaussian (LQG) control 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 to the linear quadratic Gaussian (LQG) control inputs is introduced.

In \([16]\), the authors provide a methodology to optimise the watermarking signal power, which will maximise the detection rate for a given increase in LQG control cost. In \([5]\), the authors further generalise the method and find the optimum watermarking signal in the class of Gaussian stationary processes by maximising a relaxed version of the Kullback–Leibler divergence (KLD) measure. In \([1]\), the authors design two residue signals, and the time average of them will converge to some finite values when the system is under attack, otherwise, it will be zero. It is assumed that the attacker uses a mathematical model similar to the original system to generate fake measurements, but the attacker does have any knowledge of the actual noise and the watermarking signal values. The authors have demonstrated their methodology in laboratory setup in \([17]\). The authors consider the system model with non-Gaussian process and observation noise, and design watermarking signal for such a system in \([19]\). In \([20]\), the authors design a statistical watermarking test to detect the attack on the sensors and the underlying communication channels. The problem of false data injection attacks in the presence of packet drop is studied in \([21]\) by the design of a joint Bernoulli-Gaussian watermarking. In \([18]\), the authors reduce the increase of control cost by designing a periodic watermarking signal. In \([22]\), the trade-off between the controller utility and the detectability of an attack is studied.

In this paper, we have studied the problem of the quickest attack detection, which has not been addressed directly in most of the reported work in the literature. The study on the topic of quickest change detection can be traced back several decades \([23]\). For our paper, we have followed the work presented in \([24]–[28]\). We have taken the non-Bayesian approach of change point detection where the change point or the attack point is unknown but deterministic. In \([24]\), it is assumed that the data before and after the change point need to be iid. We show in our study that the test data is iid before the attack, but after the attack, the test data does not remain iid. However, the test data is asymptotically stationary with or without the attack. The study in \([25]–[28]\) shows that under certain conditions the cumulative sum (CUSUM) test also 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) for the general non-iid case. Furthermore, the SADD asymptotically converges to the inverse of the expected value of the KLD for the non-iid case provided certain conditions are satisfied \([28]\). We have referred to the CUSUM test using the dependent distributions for the non-iid case as the optimal CUSUM test. If the CUSUM test is performed using the non-dependent distributions for the non-iid data, then we have mentioned it as a non-optimal CUSUM test. The latter may be applicable when finding the analytic form of the dependent distributions may not be feasible.

### B. Motivations and Contributions

For the safety and security of CPS, it is of paramount importance to detect the attack with minimum possible delay, thus favouring quickest sequential detection based methods. The more the attacker remains stealthy, the more damage will be caused. The watermarking based detection techniques reported in \([1]–[5], [13]\) are not specifically designed for quickest detection of attacks. Thus we will here focus on the design and analysis of the quickest sequential detection of deception attacks 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 \([29]\). We consider a linear NCS where the attacker can hijack the sensor nodes and feed fake measurement data to the estimator. The fake measurement data are assumed to be stationary and generated from a stochastic linear system. The time of the attack is unknown but deterministic in nature. 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 have reported a preliminary study on this method for the scalar case applying non-optimal CUSUM test, in \([30]\). In the current paper, we extend the work significantly by considering more generalized system models, in-depth analysis of the optimal CUSUM test for the non-iid data, and extensive numerical simulations. The proposed approach can also be applied to detect a replay attack after a few modifications as reported in \([31]\). Our main contributions are as follows.

1. 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. Since it is uncertain how long the system will be operational, probability of false alarm (PFA) may not be a practically useful metric \([32]–[33]\). We have also shown a sub-optimal sequential detection technique which will be useful where the optimal CUSUM test may not be feasible.
(ii) It is known that SADD is asymptotically inversely proportional to the expected KLD or the KLD between the joint stationary density of the innovation and watermarking signal with and without the attack under the optimal CUSUM or sub-optimal CUSUM test \cite{25, 28}. 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 demonstrate a technique to 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.

(iv) We take the joint distribution of the innovation signal and the watermarking signal to increase the KLD, unlike some of the previous works which consider only the innovation signal. An increase in KLD results in lower SADD, and thus in quicker detection.

C. Paper Organization

The organization of the remaining part of the paper is as follows. Section III 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 III. All the theorems and lemmas associated with multi-input and multi-output (MIMO) and MISO systems are provided in Section IV. The optimization technique to maximize the KLD by finding a proper watermarking signal variance is also illustrated in Section IV. We present numerical results in Section V to validate the theory. Section VI concludes the paper.

D. Notations

We have used capital bold letters, e.g., \( \mathbf{A}, \mathbf{B} \), etc. to specify matrices and small bold letters, e.g., \( \mathbf{x}, \mathbf{y} \), etc. to specify vectors, unless specified otherwise. Some special notations are given in Table I.

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

### II. 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.

#### A. System Model during Normal Operations

![Fig. 1: Schematic diagram of the system during normal operation.](image)

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

\[
x_{k+1} = Ax_k + Bu_k + w_k.
\]

Here \( x_k \in \mathbb{R}^n \) and \( u_k \in \mathbb{R}^p \) are the state and input vectors at the \( k \)-th time instant respectively, whereas \( w_k \in \mathbb{R}^m \sim \mathcal{N}(0, Q) \) is an iid process noise. \( A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{n \times p}, \) and \( Q \in \mathbb{R}^{n \times n} \). \( Q \geq 0 \). Furthermore,

\[
y_k = Cx_k + v_k
\]

where \( y_k \in \mathbb{R}^m \) is the sensor output or the observation vector at the \( k \)-th time instant. Here \( C \in \mathbb{R}^{m \times n} \), and \( v_k \in \mathbb{R}^m \sim \mathcal{N}(0, R) \) is the iid measurement noise. We assume, \( R > 0 \).

The noise vectors \( v_k \) and \( w_k \) are mutually independent, and both are independent of the initial state vector, \( x_{k_0} \). 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 sensor measurements and the input signal information, and estimates the states as follows.

\[
\hat{x}_{k|k-1} = A\hat{x}_{k-1|k-1} + Bu_{k-1}
\]

\[
\hat{x}_{k|k} = \hat{x}_{k|k-1} + K\gamma_k
\]

where \( \hat{x}_{k|k-1} = E[x_k|\Psi_{k-1}] \) and \( \hat{x}_{k|k} = E[x_k|\Psi_k] \) 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 \( K \) are given by

\[
\gamma_k = y_k - C\hat{x}_{k|k-1}
\]

\[
K = PC^T (CPC^T + R)^{-1}
\]

where \( P = E[(x_k - \hat{x}_{k|k-1})(x_k - \hat{x}_{k|k-1})^T] \) is the steady state error covariance, \( P \) is the solution to the following algebraic Riccati equation

\[
P = APA^T + Q - AP(CPC^T + R)^{-1}CPA^T.
\]

The control input \( u_k \) is generated by minimizing the following infinite horizon LQG cost

\[
J = \lim_{T \to \infty} E \left[ \frac{1}{2T+1} \sum_{k=-T}^{T} (x_k^T W x_k + u_k^T U u_k) \right]
\]
where $W \in \mathbb{R}^{n \times n}$ and $U \in \mathbb{R}^{p \times p}$ are positive definite diagonal weight matrices. The optimum input appears as a fixed gain linear control signal given by

$$u_k^* = L x_{k|k}$$

(9)

$$L = -(B^T S B + U)^{-1} B^T S A$$

(10)

where $S$ is the solution to the following algebraic Riccati equation,

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

(11)

**B. 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 can tamper with the integrity of the sensor nodes and feed undesired information to the system.

3) The attacker does not have access to the control signal or the controller.

The objective of the adversary is to cause harm to the system by replacing the true sensor measurements $y_k$ by fake observations $z_k$, and at the same time remain stealthy. The adversary can achieve his goal by jamming or overpowering the true sensor data sent over a wireless link or by hijacking the sensor nodes (man-in-the-middle attack). The adversary will also try to remain undetected as long as possible to cause maximum damage to the system. 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 $y_k$ by the fake observation $z_k$ at a deterministic but unknown time instant $k = \nu$, and keeps on injecting the fake observation for $k \geq \nu$. It is assumed that the fake observations will be generated by the following stochastic linear system

$$z_k = A_z z_{k-1} + w_{a,k-1}$$

(12)

where $z_k \in \mathbb{R}^m$, and $w_{a,k} \sim \mathcal{N}(0,Q_a)$ is the iid noise vector at the $k$-th time instant. $Q_a \in \mathbb{R}^{m \times m}$ and $Q_a \geq 0$. The attacker will try to keep the statistical properties of $z_k$, i.e., mean and variance, similar to the true observation $y_k$ to remain stealthy. Since the true measurement $y_k$ is stationary, the attacker will keep the fake measurement $z_k$ stationary by taking the initial covariance of $z_k$ as $E_{zz}(0) \equiv E[z_k z_k^T]$ to remain stealthy, where $E_{zz}(0)$ is the solution to the following Lyapunov equation,

$$E_{zz}(0) = A_a E_{zz}(0) A_a^T + Q_a.$$ 

(13)

The estimated states from the Kalman filter will take the following form when the system is under attack, i.e., $k \geq \nu$,

$$\hat{x}_{k|k-1}^F = A \hat{x}_{k-1|k-1}^F + B u_{k-1}$$

(14)

$$\hat{x}_{k|k}^F = \hat{x}_{k|k-1}^F + K \gamma_k$$

(15)

$$\gamma_k = z_k - C \hat{x}_{k|k-1}^F.$$ 

(16)

It is the same Kalman filter as given in (3)–(7) with the true observation $y_k$ replaced by the fake data $z_k$. So, the defender does not need to change anything for the Kalman filter during the attack.

An attack can make the system unstable by following the described attack model. For illustration, the true and estimated states of System-A 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 Appendix I. The system becomes unstable soon after the attack.

![Fig. 3: True and estimated states of System-A.](image)

**III. PHYSICAL WATERMARKING BASED DEFENCE MECHANISM AND DELAY IN DETECTION**

This section proposes the physical-watermarking-based sequential attack detection scheme and discusses about the delay in the detection process. We use hypothesis testing to detect the attack. There are two different hypotheses to choose from,

- $H_0$: No attack. Estimator receives the true observation $y_k$
- $H_1$: Attack. Estimator receives a fake observation $z_k$.

The innovation signals ((16) 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 $\gamma_k^f$ are denoted as $f_{\gamma_k}(\gamma_k)$ and $f_{\gamma_k^f}(\gamma_k)$ respectively, where $\gamma_k = \gamma_k^f$ before attack, and $\gamma_k = \gamma_k^f$ after attack. Both the distributions $f_{\gamma_k}(\gamma_k)$ and $f_{\gamma_k^f}(\gamma_k)$ are stationary in nature. The probability of attack detection will increase if the KLD i.e., $D(f_{\gamma_k}, f_{\gamma_k^f})$, between the two distributions $f_{\gamma_k}(\gamma_k)$ and $f_{\gamma_k^f}(\gamma_k)$ under $H_1$ and $H_0$ increases [27].

$$D(f_{\gamma_k}, f_{\gamma_k^f}) = \int_{\mathbb{R}^m} f_{\gamma_k}(\gamma_k) \log \frac{f_{\gamma_k}(\gamma_k)}{f_{\gamma_k^f}(\gamma_k)} d\gamma_k.$$ 

(17)

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.
A. Physical Watermarking

A well-adopted technique to detect attacks on the control system is to add a watermarking signal, as described above [1], [5]. The control designer thus adds a random watermarking signal $e_k$ to the optimal LQG control input $u_k^*$, see [15]. 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.

$$ u_k = u_k^* + e_k $$

where $u_k^*$ is the optimal input [9], $e_k \sim N(0, \Sigma_e)$ is an iid process, and $\Sigma_e \geq 0$, and possibly non-diagonal matrix. In the literature, $e_k$ 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 $e_k$ 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 $y_k$, which is correlated to $e_{k-1}$, is replaced by $z_k$, which is independent of $e_{k-1}$, even if the attacker knows the statistics of $e_k$.

Detection of the attack as early as possible is of utmost importance to reduce the damage. The optimal Neyman-Pearson (NP) test [5] and the asymptotic test [1] 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 [27] 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 $e_{k-1}$, and $\tilde{\gamma}_k$ and $e_{k-1}$ for the test. We show the simulation results in the Section V 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 [19] and [20], respectively.

$$ \gamma_k = y_k - C\tilde{x}_{k|k-1} $$

$$ \gamma_k = y_k - C\hat{x}_{k|k-1} $$

$$ \tilde{\gamma}_k = z_k - C\tilde{x}_{k|k-1} $$

$$ \tilde{\gamma}_k = z_k - C\hat{x}_{k|k-1} $$

It is evident from [19] and [20] 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.

B. Detection Delay

We use the delay in the attack detection as the metric to measure the performance of the defense 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 [27]. We start this section by introducing the definitions of relevant terms as follows.

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

$$ ADD \triangleq E_\nu [T_{H_1} - \nu | T_{H_1} > \nu] $$

where $E_\nu[\cdot]$ is the expectation taken with respect to the PDF under attack. Here $\nu$ is the attack starting point in time which is assumed to be unknown but deterministic in nature, whereas $T_{H_1}$ is the attack starting point detected by a hypothesis testing algorithm.

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

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

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

$$ ARL \triangleq E_\nu [T_{H_1}] $$

where $E_\nu[\cdot]$ 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 [27]. We can only find a procedure that will minimize the worst-case ADD for any $\nu$, i.e., ADD, for a fixed threshold on ARL. As per the theory presented in [27], CUSUM is one of such procedures. The CUSUM procedure is asymptotically minimax in the sense of minimizing the SADD for all $\nu > 0$, as $ARL_h \to \infty$, and the minimum SADD is

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

where $I$ is a finite positive real number, $ARL_h$ is the threshold on ARL, $ARL \geq ARL_h$, provided the following three conditions are satisfied [27]:

i) \[ \frac{1}{n} \sum_{k=n+1}^{n+\nu} \frac{P_{\nu}}{M} \rightarrow I, \]

ii) \[ \sup_{0 \leq \nu < \infty} \sup_{\nu \leq \nu < \infty} \max_{0 \leq M} \lambda_{\nu+n} \geq 1 + \epsilon \]

iii) \[ \sup_{0 \leq \nu < \infty} \sup_{\nu \leq \nu < \infty} \max_{0 \leq M} \lambda_{\nu+n} \geq 1 - \epsilon \]

where $P_{\nu}$ indicates the probability after the change and $M$ is a positive integer variable. Here $\Psi_\nu$ is the set of all observations up until the change point $\nu$. The variable $\lambda_{\nu+n}$ is defined as

$$ \lambda_{\nu+n} \triangleq \sum_{k=\nu+1}^{n+\nu} \log \frac{f_{\nu,k}(X_k|\{X\}_1^{k-1})}{f_{\infty,k}(X_k|\{X\}_1^{k-1})} $$

where $X_k$ is the observation at the $k$-th time instant and $\{X\}_1^{k-1} = \{X_i : 1 \leq i \leq k - 1\}$. In [28], $f_{\nu,k}(\cdot|\cdot)$ 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,

$$ \lambda_{\nu+n} = \sum_{k=\nu+1}^{n+\nu} \log \frac{f_{\tilde{\gamma}_k,e_{k-1}}(\tilde{\gamma}_k,e_{k-1} | \{\gamma\}_1^{k-1}, \{e\}_1^{k-2})}{f_{\gamma_k,e_{k-1}}(\tilde{\gamma}_k,e_{k-1} | \{\gamma\}_1^{k-1}, \{e\}_1^{k-2})} $$
where \( f_{\gamma_k, e_{k-1}}(\cdot) \) and \( f_{\gamma_k, e_{k-1}}(\cdot) \) are the joint dependent 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. \( \{\gamma_i\}_{i=1}^{k-1} = \{\gamma_i : 1 \leq i < \nu\} \cup \{\tilde{\gamma}_i : \nu \leq i \leq k-1\} \). The data \((\gamma_k, \tilde{\gamma}_k)\) and \((e_{k-1})\) 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 \( f_{\gamma_k, e_{k-1}}(\cdot) \) and \( f_{\gamma_k, e_{k-1}}(\cdot) \) as \( n \to \infty \). In other words,

\[
I \to \frac{1}{n} \sum_{k=\nu+1}^{n+\nu} \log f_{\gamma_k, e_{k-1}}(\gamma_k, e_{k-1} | \{\gamma_i\}_{i=1}^{k-1}, \{e_i\}_{i=1}^{k-2}),
\]

as \( n \to \infty \), which converges to the following form,

\[
E \left[ \log f_{\gamma_k, e_{k-1}}(\gamma_k, e_{k-1} | \{\gamma_i\}_{i=1}^{k-1}, \{e_i\}_{i=1}^{k-2}) \right] = E \left[ D \left( f_{\gamma_k, e_{k-1}}(\gamma_k, e_{k-1}, | \{\gamma_i\}_{i=1}^{k-1}, \{e_i\}_{i=1}^{k-2}) \right) \right].
\]

Here, the expectation is taken over the joint distribution of \( \{\gamma_i\}_{i=1}^{k-1}, \{e_i\}_{i=1}^{k-2} \).

### C. Optimal and Sub-optimal CUSUM Tests

The following CUSUM test will minimize the SADD asymptotically,

\[
gd_k = \max \left( \log f_{\gamma_k, e_{k-1}}(\gamma_k, e_{k-1} | \{\gamma_i\}_{i=1}^{k-1}, \{e_i\}_{i=1}^{k-2}), f_{\gamma_k, e_{k-1}}(\gamma_k, e_{k-1}) \right)
\]

where \( \gamma_k = \gamma_k \) before attack, and \( \gamma_k = \tilde{\gamma}_k \) after attack, and

\[
SADD^* \to \log(ARL_h) \to \infty.
\]

Since before the attack the innovation signal \( \gamma_k \) and the watermarking signal \( e_{k-1} \) both are iids, and also uncorrelated to each other, the non-dependent distribution is used in the denominator of (31). The controller decides on hypothesis \( H_0 \) or \( H_1 \) based on the following test,

- \( H_0 \): Selected, when \( \log(ARL_h) \)
- \( H_1 \): Selected, when \( \log(ARL_h) \)

For certain cases, the closed-form expressions for the dependent 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 non-dependent distributions for sequential attack detection,

\[
g_k = \max \left( \log f_{\gamma_k, e_{k-1}}(\gamma_k, e_{k-1}), f_{\gamma_k, e_{k-1}}(\gamma_k, e_{k-1}) \right)
\]

### IV. 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 theorems for the general MIMO systems in Subsection IV-A, and then simplify the theorems for the MISO systems in Subsection IV-B to acquire better structural understanding. The technique to optimize the \( \Sigma_e \) to achieve minimum SADD for a given upper bound on the \( \Delta LQG \) is illustrated in Subsection IV-C.

#### A. Multiple Input Multiple Output Systems

**Theorem 1.** The optimal CUSUM test to detect the deception attack given by (12) will take the following form,

\[
gd_k = \max \left( \log f_{\gamma_k}(\gamma_k), f_{\gamma_k}(\gamma_k) \right)
\]

where \( \gamma_k = \gamma_k \) before attack, and \( \gamma_k = \tilde{\gamma}_k \) after attack,

\[
\{\gamma_k | \{\gamma_i\}_{i=1}^{k-1}, \{e_i\}_{i=1}^{k-2}\}
\]

\[
\sim N \left( \mu_{\gamma_k | \{\gamma_i\}_{i=1}^{k-1}, \{e_i\}_{i=1}^{k-2}}, \Sigma_{\gamma_k | \{\gamma_i\}_{i=1}^{k-1}, \{e_i\}_{i=1}^{k-2}} \right)
\]

\[
\mu_{\gamma_k | \{\gamma_i\}_{i=1}^{k-1}, \{e_i\}_{i=1}^{k-2}} = \begin{cases} \mathbf{A}_k \mathbf{z}_{k-1} - \mathbf{C} (\mathbf{A} + \mathbf{BL}) \tilde{x}^F_{k-1,k-1} - \mathbf{C} \mathbf{B} \mathbf{e}_{k-1}, & k \geq \nu \cr \mathbf{A}_k \mathbf{y}_{k-1} - \mathbf{C} (\mathbf{A} + \mathbf{BL}) \tilde{x}^F_{k-1,k-1} - \mathbf{C} \mathbf{e}_{k-1}, & k < \nu \end{cases}
\]

\[
\Sigma_{\gamma_k | \{\gamma_i\}_{i=1}^{k-1}, \{e_i\}_{i=1}^{k-2}} = \mathbf{Q}_e, \quad \text{and}
\]

\[
\gamma_k \sim N \left( 0, \Sigma_{\gamma_k} \right),
\]

\[
\Sigma_{\gamma_k} = \mathbf{C} \Sigma \mathbf{D}^T + \mathbf{R}.
\]

**Proof.** The proof of Theorem [A] is provided in Appendix [A].

**Remark 1.** The likelihood ratio in (35) will be evaluated using the innovation signal \( \gamma_k \) from the Kalman filter, \( \gamma_k = \gamma_k \) if \( k < \nu \), and it will change automatically to \( \tilde{\gamma}_k = \gamma_k \) if \( k \geq \nu \) without any intervention from the defender. Similarly, \( \mathbf{y}_k \) and \( \mathbf{x}_{k-1,k-1} \) will change to \( \mathbf{y}_k \) and \( \mathbf{x}_{k-1,k-1} \) respectively, after the attack, as given in (36). However, the attacker plays an active role by replacing the true observation \( \mathbf{y}_k \) by the fake data \( \mathbf{z}_k \) at \( k \geq \nu \).

**Remark 2.** The optimal CUSUM test utilising the dependent distributions of the innovation signals before and after an attack is performed employing Theorem [A]. The innovation signal
The sub-optimal CUSUM test using the non-conditional distributions to detect the deception attack given by (12) will take the following form,

\[ g_k = \max \left( 0, g_{k-1} + \log \frac{f_{\gamma_k,e_{k-1}}(\gamma_k,e_{k-1})}{f_{\gamma_k,e_{k-1}}(\gamma_k,e_{k-1})} \right) , \]  

(39)

where \( \gamma_k = \gamma_k \) before attack, and \( \gamma_k = \gamma_k \) after attack, \( \gamma_{r,k} = \gamma_T^{T_k} e_{k-1}^{T} \sim N(0, \Sigma_{\gamma_k}) \), \( \gamma_{c,k} = \gamma_T^{T_k} e_{k-1}^{T} \sim N(0, \Sigma_{\gamma_c}) \),

(40)

where \( \Sigma_{\gamma_c} = \begin{bmatrix} \Sigma_{\gamma} & 0_{m \times p} \\ 0_{p \times m} & \Sigma_e \end{bmatrix} \) and \( \Sigma_{\gamma_r} = \begin{bmatrix} \Sigma_{\gamma} & -CB \Sigma_e \\ -C \Sigma_e B^T C^T & \Sigma_e \end{bmatrix} \).

(41)

**Remark 4.** Both the test statistics \( g_d \) and \( g_e \) will be close to zero during the normal operation, and they will gradually increase after the attack at every time step.

**Remark 5.** For the sub-optimal CUSUM test, the non-dependent and asymptotically stationary distributions of \( \gamma_k \) and \( \gamma_k \) are used. Such a test can be applied when designing the optimal CUSUM test is not feasible, e.g., replay attack detection as discussed in [37]. Also, for the optimal CUSUM test, the dependent mean needs to be evaluated at every time step, which increases the computational complexity compared to the sub-optimal CUSUM test.

**Lemma 1.** The covariance matrix \( \Sigma_{\gamma} \) of the innovation signal \( \gamma \) after the attack will take the following form,

\[ \Sigma_{\gamma} = E_{zz}(0) - C(A + BL)E_{xz}(-1) - [C(A + BL)]^T E_{zz}(0) + C(A + BL)\Sigma_{\gamma_f}(A + BL)^T C^T + C(A + BL)\Sigma_{\gamma_r}(A + BL)^T C^T , \]

(42)

where \( E_{zz}(0) = \sum_{i=0}^{\infty} A_i^T K A_i^{i+1} E_{zz}(0) \)

(43)

and \( E_{zz}(0) = E \left[ z_k z_k^T \right] \), \( \Sigma_{\gamma_f} \) and \( \Sigma_{\gamma_r} \) are the solutions to the following Lyapunov equations,

\[ \begin{align*}
A \Sigma_{\gamma_f} & = \Sigma_{\gamma_f} A^T - \Sigma_{\gamma_f} + KE_{zz}(0) K^T + AE_{zz}(-1) K^T \quad \text{and} \\
A \Sigma_{\gamma_r} & = \Sigma_{\gamma_r} A^T - \Sigma_{\gamma_r} + (I_n - KC) B \Sigma_{e} B^T (I_n - KC)^T \quad \text{if } \Sigma_{\gamma} \neq 0 .
\end{align*} \]

(44)

**Remark 6.** Since \( \Sigma \) is assumed to be strictly stable, the Lyapunov equations of (42) and (44) will have unique solutions. If \( \Sigma \) and \( A \) are not diagonalizable, then \( E_{zz}(-1) \) can be evaluated numerically by taking a large number of terms for the summation of (43) before the rest of the terms become negligible.

**Remark 7.** The attacker’s system parameters \( A_a \) and \( Q_a \) can be estimated from the observations.

**Corollary 1.2.** With the assumption that \( A \) and \( A_a \) are diagonalizable, \( E_{xz}(-1) \) will take the following form

\[ E_{xz}(-1) = U_A T_A U^{-1}_{a} A_a E_{zz}(0) . \]

(46)

Here \( U_A \) is the eigenvector matrix of \( A \), see (47). \( \Sigma_A = \text{diag} \{ \lambda_{A,1}, \lambda_{A,2}, \ldots \} \) is the eigenvalue matrix of \( A \) with the eigenvalues on its main diagonal. \( U_a \) is the eigenvector matrix of \( A_a \), see (48). \( \Sigma_a = \text{diag} \{ \lambda_{a,1}, \lambda_{a,2}, \ldots \} \) is the eigenvalue matrix of \( A_a \) with the eigenvalues on its main diagonal.

\[ \begin{align*}
A & = U_A \Sigma_A U^{-1}_A , \\
A_a & = U_a \Sigma_a U^{-1}_a .
\end{align*} \]

(47)

The \( ij \)-th element of the \( T_a \) matrix is as follows

\[ [T_a]_{ij} = \frac{[T]_{ij}}{1 - \lambda_{A,1} \lambda_{a,j}} , \]

(49)

and \( T = U^{-1}_a K U_a \).

(50)

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

**Remark 8.** Corollary 1.2 provides a way to derive the value of \( E_{xz}(-1) \) analytically, provided \( A \) and \( A_a \) are diagonalizable. \( E_{xz}(-1) \) is used to evaluate \( \Sigma_{\gamma} \).

**Theorem 2.** The expected KLD under the optimal CUSUM test \( E \left[ D \left( f_{\gamma_k}, f_{\gamma_k} | \gamma_{1,k}^{k-1}, \{e_{1,k}^{k-1}\} \right) \right] \), and the KLD under the sub-optimal CUSUM test \( D \left( f_{\gamma_k,e_{k-1}}, f_{\gamma_k,e_{k-1}} \right) \) will be as follows,

\[ \begin{align*}
\mathbb{E} & \left[ D \left( f_{\gamma_k}, f_{\gamma_k} | \gamma_{1,k}^{k-1}, \{e_{1,k}^{k-1}\} \right) \right] \\
& = \frac{1}{2} \left\{ \text{tr} \left( \Sigma^{-1} \Sigma_{\gamma} \right) - m - \log \left| \frac{Q_a}{\Sigma_{\gamma}} \right| \right\} , \quad \text{and} \\
D & \left( f_{\gamma_k,e_{k-1}}, f_{\gamma_k,e_{k-1}} \right) \\
& = \frac{1}{2} \left\{ \text{tr} \left( \Sigma^{-1} \Sigma_{\gamma} \right) - m - \log \left| \Sigma_{\gamma} - CB \Sigma_e B^T C^T \right| \right\} ,
\end{align*} \]

(51)

(52)

**Proof.** The proof of Theorem 2 is provided in the Appendix E.
Corollary 2.1. The difference between the expected KLD and the KLD is log \( \frac{|\Sigma_e - \Sigma_{\tilde e}|}{|Q_e|} \), which corresponds to the optimality gap between the optimal and sub-optimal CUSUM tests. From (17), exploiting suitable independence properties of the involved random processes, it can be shown that \( \Sigma_e - \Sigma_{\tilde e} \geq Q_e \). By eigenvalue comparison of the positive semidefinite matrices \( \Sigma_e - \Sigma_{\tilde e} \) and \( Q_e \), we can say \( |\Sigma_e - \Sigma_{\tilde e}| \geq |Q_e| \), which ensures the optimality gap is positive.

Proof. The proof simply follows by subtracting (52) from (51).

Remark 9. The expected KLD and the KLD under the optimal and sub-optimal test, respectively, are mostly dependent on the non-dependent variances of the innovation signals \( \Sigma_e \) and \( \Sigma_{\tilde e} \) before and after an attack. They also depend on a few system and noise parameters.

Remark 10. Instead of taking the joint distribution of the innovation signal and the watermarking signal, if the optimal CUSUM test is performed using the dependent distribution of the innovations signal only, then the expected KLD will take the form of (53). While a detailed proof cannot be accommodated due to space constraints, here we use simple intuitive arguments to explain why the expected KLD of (53) reduces compared to the optimal KLD using the joint conditional distribution of the innovation signal and the watermarking signal (57). 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 increasing the KLD overall. The increase in KLD results in quicker attack detection on average due to (24). Equation (53) can be derived following the similar steps given in the Appendix A. However, the detailed proof has been omitted due to the space constraints.

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 \( \Sigma_e \) as follows,

\[
\Delta LQG = \text{tr} (H \Sigma_e)
\]

where \( H = B^T \Sigma_L B + U \) (57) and \( \Sigma_L \) is the solution to the Lyapunov equation

\[
(A + BL)^T \Sigma_L (A + BL) - \Sigma_L + L^T UL + W = 0.
\] (59)

Proof. The theorem can be proved easily using the Theorem 2 from [5], considering the iid watermarking as a special case of the hidden Markov model (HMM).

Remark 11. Since the closed loop system \( (A + BL) \) is stable, the Lyapunov equation of (59) will have a unique solution.

Remark 12. 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 \( \Sigma_e \) of the added watermarking. The matrix \( H \) in (57) is dependent on the plant and controller parameters. Since the plant and the controller are assumed to be time-invariant, \( H \) 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, \( \Sigma_e \), of the watermarking signal.

B. 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 1. The following attack model is assumed for the MISO system, which is a special case of the stochastic linear attack model given in (12).

\[
E[\tilde{z}_k^2] = \sigma_z^2, \quad \text{and} \quad E[z_k z_k - k_o] = \rho k o \sigma_z^2, \quad \rho < 1.
\] (60)

Therefore, \( A_d = \rho \), and \( Q_d = (1 - \rho^2) \sigma_z^2 \).

Lemma 2. For a MISO system, the expected KLD \( E \left[ \ln \left( f_{\tilde{z}_k, f_{z_k}} \right) \right] \) under the optimal CUSUM test, and the KLD \( D \left( f_{\tilde{z}_k, e_k, -1} \right) \) under the sub-optimal CUSUM test will be as follows,

\[
D(f_{\tilde{z}_k, e_k, -1}) = \frac{1}{2} \left( \frac{\sigma_{e_k}^2}{\sigma_{\tilde{e}_k}^2} - 1 - \log \left( \frac{1}{\rho^2} \right) \frac{\sigma_{e_k}^2}{\sigma_{\tilde{e}_k}^2} \right)
\] (61)

where the attack model is given by (60). \( \sigma_{e_k}^2 \) and \( \sigma_{\tilde{e}_k}^2 \) are the scalar variances of the innovation signals in \( e_k \) and \( \tilde{e}_k \) before and after the attack, respectively. Hence,

\[
\sigma_{e_k}^2 = CPC^T + R, \quad \text{and} \quad \sigma_{\tilde{e}_k}^2 = M_e \sigma_{e_k}^2 + \text{tr} (M_e \Sigma_e)
\] (62)

where \( R \) and \( M_z \) are scalar quantities. \( M_z \) and \( M_e \) will take the following forms,

\[
M_z = 1 - 2C (A + BL) (I_n - \rho A)^{-1} K \rho + C (A + BL) \Sigma_{x, \rho} (A + BL)^T C^T, \quad \text{and}
\]

\[
M_e = B^T (I_n - KC)^T \Sigma_{e, \rho} (I_n - KC) B + B^T C^T CB
\] (66)
Proof. (61) and (62) can be derived directly by replacing in Appendix F.

Furthermore, the KLD, but at the same time, it also increases the control system designer can not do much to avoid this situation. On the other hand, the control system designer can choose the attacker signal power \( \sigma^2 \) for the sub-optimal case, so that the KLD will always increase with the attack signal power. However, under the optimal test, the control system designer cannot do the optimal test does not depend on the watermarking signal power. In fact, the attacker can modify \( \sigma^2 \) to \( \sigma^2 \) to reduce the KLD which in turn reduces the probability of detection. On the other hand, the control system designer can choose \( \text{tr} \left( [M_e - B^T C T C] \Sigma_e \right) \geq \sigma^2 \) 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 \( \sigma^2 \).

C. Optimum Watermarking Signal for MISO systems

By increasing the watermarking power \( \Sigma_e \), 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 \( \Sigma_e \), say \( \Sigma^*_e \), which will maximize the KLD subject to an upper bound on \( \Delta LQG \). The optimization problem is formulated as follows,

\[
\begin{align*}
\max_{\Sigma_e} E \left[ D \left( f_{\gamma_k}, f_{\gamma_1} \{\gamma\}^k-1 \right) \right] \quad \text{or} \\
\max_{\Sigma_e} D \left( f_{\gamma_k, e_{k-1}}, f_{\gamma_1, e_{k-1}} \right) \\
s.t. \Delta LQG \leq J \\
\Sigma_e \geq 0
\end{align*}
\]

where \( J \) is a user choice. The positive semi-definite \( \Sigma_e \) matrix can be decomposed by the eigenvalue decomposition as

\[
\Sigma_e = V_e \Lambda_e V_e^T ,
\]

where \( V_e \) is the orthonormal eigenvector matrix and \( \Lambda_e \) is the diagonal eigenvalue matrix. If we assume that \( V_e \) is known a priori, then we only need to find the optimum \( \Lambda_e \) which is a diagonal matrix.

Theorem 4. The optimum diagonal \( \Lambda_e \) 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 F.

In the light of Theorem 4, we search for the optimum \( \Sigma_e \) in the class of rank one positive semi-definite matrices of the following form

\[
\Sigma_e = \lambda_e v_e v_e^T ,
\]

where \( \lambda_e \) is the non-zero eigenvalue and \( v_e \) is the corresponding eigenvector. We modify (58) to represent it in the following form

\[
\Sigma_e = v_\lambda^2 v_\lambda' ,
\]

where \( v_\lambda \) is the eigenvector associated with \( \lambda_e \). Finally, the optimization problem becomes,

\[
\begin{align*}
\max_{v_\lambda} E \left[ D \left( f_{\gamma_k}, f_{\gamma_1} \{\gamma\}^k-1 \right) \right] \quad \text{or} \\
\max_{v_\lambda} D \left( f_{\gamma_k, e_{k-1}}, f_{\gamma_1, e_{k-1}} \right) \\
s.t. \Delta LQG \leq J.
\end{align*}
\]
The optimization problem can be solved using different methods such as the sequential quadratic programming (SQP) [34], the interior point method [35], etc. We have also provided a simple gradient descent based algorithm to solve the optimization problem (75)-(76) in Appendix-H.

The cost function under the optimal CUSUM test can be simplified. Maximization of

\[
E \left[ D \left( \tilde{\gamma}_k, f_{\gamma_k} \right) \left| \left\{ \tilde{\gamma}_1^{k-1}, \{e_i\}_{1}^{k-1} \right\} \right. \right]
\]

with respect to \( \nu_\lambda \) is the same as maximizing the following function with respect to \( \nu_\lambda \).

\[
\nu_\lambda^T H_{KLD} \nu_\lambda
\]

where

\[
H_{KLD} = B^T (I_n - KC)^T L_e (I_n - KC) B + B^T C^T CB
\]

(77)

and \( L_e \) is the solution to the Lyapunov equation

\[
A^T L_e A - L_e + (A + BL)^T C^T C (A + BL) = 0
\]

(78)

Since the matrix \( A \) is assumed to be strictly stable, the Lyapunov equation of (78) will have unique solution. The derivations are provided in Appendix-H. (77) and (78) can be simplified for the system with relative degree higher than one, since \( CB = 0 \).

V. 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-I. System-B is a linearized minimum phase quadruple tank system which is used previously to test the deception attack detection schemes in the literature [36]. Only the level sensor gains are altered to make the magnitude of the product \( CB \) numerically significant.

A. Tradeoff between SADD and \( \Delta LQG \) 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 (35). We plot the derived SADD using the theory developed in this paper, and the estimated SADD from Monte-Carlo (MC) simulation, where \( \Sigma_e \) is assumed to be diagonal and both the watermarking signals have equal power. An increase in LQG cost results in quicker detection.

B. 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 detection 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 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

\[
\frac{\Delta LQG_2 - \Delta LQG_1}{\Delta LQG_1} \times 100\%
\]

C. 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. 5. We assume, \( \Delta LQG = 100 \), and \( \Sigma_e \) 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 \( \sigma_z^2 \).

D. Optimum vs non-optimum \( \Sigma_e \)

We optimize the \( \Sigma_e \) under the optimal test using the method in Subsection IV-C. Figure 7 shows the SADD vs \( \Delta LQG \) plots...
using the optimized $\Sigma_e$ and a diagonal $\Sigma_e$ 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 $\Sigma_e$ and non-optimized $\Sigma_e$ in the figure. It is evident that optimizing $\Sigma_e$ 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 $\Delta LQG_2$.

**E. Optimal vs sub-optimal CUSUM**

Figure 8 illustrates the advantage of performing the optimal CUSUM test with dependent PDFs over the sub-optimal CUSUM test using the non-dependent PDFs for System-A. For both the plots, optimum $\Sigma_e$ 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.

**F. Comparison with optimal NP detector**

We have compared the optimal CUSUM test results with the optimal NP detector based method reported in [5], [16]. The watermarking signal is taken to be iid, and the $\Sigma_e$ is optimized for both the cases. In [5], the optimal NP detector rejects the $H_0$ hypothesis in favour of $H_1$ if

$$g_{NP,k}(\gamma_k, e_{k-1}, \cdots) = \gamma_k^T \Sigma_e^{-1} \gamma_k - (\gamma_k - \mu_{NP,k})^T (\Sigma_e + \Sigma_f)^{-1} (\gamma_k - \mu_{NP,k}) \geq \eta$$

(79)

where $\mu_{NP,k} = -C \sum_{i=-\infty}^k \mathcal{A}^{k-i} \mathbf{B} e_i$, $\Sigma_f = C \mathcal{L}_f C^T$, and $\mathcal{L}_f = \mathcal{A} \mathcal{L}_f \mathcal{A}^T + \mathbf{B} \Sigma_e \mathbf{B}^T$. (80)

The threshold $\eta$ is estimated by simulation from

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

(83)

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 [37], [38]. For the method in [5], the ADD is estimated as

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

(84)

Figure 9 illustrates how the test statistics $g_{k}$ and $g_{NP,k}$ vary with time $k$ under the optimal CUSUM [35] 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. Figure 10 shows the tradeoff between the ADD and the increase in $\Delta LQG$ for System-A under the optimal CUSUM test and the method reported in [5]. We plot the derived SADD using the theory developed in this paper, the estimated SADD applying the optimal CUSUM test on the simulated data, and the estimated ADD applying the test reported in [5] on the simulated data. It is clear from the figure that we can achieve lower ADD for the same LQG loss with the method proposed in this paper compared to the one reported in [5].

**VI. 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 optimalCUSUM test with the optimal NP test to detect the deception attack and found the optimal CUSUM test to be quicker. In the future, the sequential attack detection scheme can be extended to detect other kinds of attacks as well. The problem of attack detection can also be formulated as a dynamic two-player game between the control system designer and the attacker. This is a topic for future research.

APPENDIX A
Proof of Theorem [1]

Under the optimalCUSUM test, the likelihood ratio from (31) can be simplified using the chain rule of probability as

\[
\frac{f_{\gamma k}(\gamma k | \{\gamma\}_{k}^{k-1}, \{e\}_{k}^{k-1})}{f_{\gamma k}(\gamma k | \{e\}_{k}^{k-1})} f_{e_{k-1}}(e_{k-1})
\]

(85)

[\{e\} is iid and stationary, and \gamma and \{e\} are uncorrelated]

\[=\frac{f_{\gamma k}(\gamma k | \{\gamma\}_{k}^{k-1}, \{e\}_{k}^{k-1})}{f_{\gamma k}(\gamma k)} [\text{provided } f_{e_{k-1}}(e_{k-1}) \neq 0],\]

(86)

where \gamma k = \gamma k before the attack, and \gamma k = \gamma k after the attack. The conditional mean (\mu_{\gamma k}^{\{\gamma\}_{k}^{k-1}, \{e\}_{k}^{k-1}}) and covariance (\Sigma_{\gamma k}^{\{\gamma\}_{k}^{k-1}, \{e\}_{k}^{k-1}}) of \gamma k are derived as follows.

The innovation signal under attack from (20) can be written as (87) after replacing \{z\} by [12],

\[
\gamma k = w_{a,k-1} + A_{a}z_{k-1} - C (A + BL) \hat{x}_{k-1|k-1} - CB e_{k-1}.
\]

(87)

Applying (14), (18), (4) in (15) we can write,

\[
\hat{x}_{k|k}^{F} = (A + BL) \hat{x}_{k-1|k-1} + Be_{k-1} + K \gamma k - 1.
\]

(88)

Using (88) recursively we get,

\[
\hat{x}_{k|k}^{F} = (A + BL)^{k-1} \hat{x}_{1|1}
+ \sum_{i=1}^{k-1} (A + BL)^{i-1} (Be_{k-i-1} + K \gamma k - i)
\]

where \gamma k = \gamma k for k < \nu, \gamma k = \gamma k otherwise. (89)

Applying (20) and (89) in (87) we get,

\[
\gamma k = w_{a,k-1} + (A_{a}C - C (A + BL)) (A + BL)^{k-1} \hat{x}_{1|1}
+ \sum_{i=1}^{k-2} (A + BL)^{i-1} Be_{k-i-1} + \sum_{i=2}^{k-2} (A + BL)^{i-1} K \gamma k - i
\]

- CB e_{k-1} + (A_{a} - C (A + BL) K) \gamma k - 1.

(90)

Since we have assumed that the system started at k = -\infty, and (A + BL) is strictly stable, we can say (A + BL)^{k-2} \approx 0, and (91) will take the following form

\[
\gamma k = w_{a,k-1} + (A_{a}C - C (A + BL))
(\sum_{i=1}^{k-2} (A + BL)^{i-1} Be_{k-i-1} + \sum_{i=2}^{k-2} (A + BL)^{i-1} K \gamma k - i)
\]

- CB e_{k-1} + (A_{a} - C (A + BL) K) \gamma k - 1.

(91)

Therefore,

\[
\mu_{\gamma k}^{\{\gamma\}_{k}^{k-1}, \{e\}_{k}^{k-1}} = E \left[ \gamma k | z_{k-1}, \hat{x}_{k-1|k-1}, e_{k-1} \right] = A_{a}z_{k-1} - C (A + BL) \hat{x}_{k-1|k-1} - CB e_{k-1}, \text{ and}
\]

\[
\Sigma_{\gamma k}^{\{\gamma\}_{k}^{k-1}, \{e\}_{k}^{k-1}} = cov \left[ \gamma k | z_{k-1}, \hat{x}_{k-1|k-1}, e_{k-1} \right] = Q_{a}.
\]

(92)

(93)

Furthermore, using (19) we obtain E [\gamma k] = 0 and

\[
\gamma k = y_{k} - Cx_{k|k-1} - C (x_{k} - x_{k|k-1}) + v_{k}, \text{ and}
\]

\[
\Sigma_{\gamma} = E [\gamma k \gamma k^{T}] = CPC^{T} + R.
\]

(94)

APPENDIX B
Proof of Corollary [1,3]
The covariance matrix (E [\gamma k e_{k-1}^{T}]) between \gamma k and e_{k-1} is evaluated as,

\[
E [\gamma k e_{k-1}^{T}] = E [-CB e_{k-1} e_{k-1}^{T}] = -CB \Sigma_{e},
\]

(95)

since e_{k-1} is uncorrelated with z_{k} and \hat{x}_{k-1|k-1}.

APPENDIX C
Proof of Lemma [1]
The variance of the innovation signal (\Sigma_{\gamma}) when the system is under attack is derived in this section. Using (20), and applying the knowledge that e_{k-1} is uncorrelated with z_{k} and \hat{x}_{k-1|k-1}, we get the following expression of \Sigma_{\gamma},

\[
\Sigma_{\gamma} = E [\gamma k \gamma k^{T}] = E [z_{k} z_{k}^{T}] - C (A + BL) E \left[ \hat{x}_{k-1|k-1} z_{k}^{T} \right]
\]

- \left( C (A + BL) E \left[ \hat{x}_{k-1|k-1} z_{k}^{T} \right] \right)^{T} + CB \Sigma_{e} B^{T} C^{T}
\]

+ C (A + BL) E \left[ \hat{x}_{k-1|k-1} \left( \hat{x}_{k-1|k-1}^{T} \right) \right] (A + BL)^{T} C^{T}.

(96)
We first derive the expressions of $E\left[ x_{k-1|k-1}^T z_k^T \right]$ and $E\left[ x_{k-1|k-1}^F \right]$ and then use them to get the final expression of $\Sigma_z$ (106). $E\left[ x_{k-1|k-1}^T z_k^T \right]$ is calculated using (14)-(16) and (18) as follows. First note that

$$\hat{x}_{k-1|k-1} = Kz_k + AX_k - 1 z_{k-1} + (I_n - KC) B e_{k-2},$$

where $A = (I_n - KC)(A + BL)$.

(97)

We define $E_{xz}(-k_0) \triangleq E\left[ x_{k-1|k-1}^T z_k^T \right]$, and use them to get the final expression of $\Sigma_z$ (106). $E\left[ x_{k-1|k-1}^T z_k^T \right]$ is calculated using (14)-(16) and (18) as follows. First note that

$$\hat{x}_{k-1|k-1} = Kz_k + AX_k - 1 z_{k-1} + (I_n - KC) B e_{k-2},$$

where $e_{k-1|k-1}$ and $z_k$ are uncorrelated, and $E_{xz}(-k_0)$ is evaluated as follows.

$$E_{xz}(-k_0) = E_{xz} (k_0) = E \left[ z_k z_k^T \right],$$

$$E_{xz}(-1) = E \left[ A a x_{k-1} z_{k-1} + w_{a,k} x_{k-1} \right] = A a E_{xz} (0),$$

because $w_{a,k}$ and $z_k$ are uncorrelated. Similarly,

$$E_{xz}(-2) = A a E_{xz} (-1) = A a E_{xz} (0),$$

and $E_{xz}(-k_0) = A a E_{xz} (0)$. (99)

The system matrix $A a$ is assumed to be strictly stable because the attacker will always try to generate false observations which are bounded and will mimic the true observations to remain stealthy. For a strictly stable $A a$,

$$A a^k \rightarrow 0, \text{ as } k_0 \rightarrow \infty.$$

Therefore, $E_{xz} (-k_0) \rightarrow 0, \text{ as } k \rightarrow \infty.$

(100)

Using (98) and (99), we can write the expression of $E_{xz}(-1)$ as

$$E_{xz}(-1) = K E_{xz} (-1) + A E_{xz} (-2)$$

$$= K A a E_{xz} (0) + A (K E_{xz} (-2) + A E_{xz} (-3))$$

[after replacing $E_{xz}(-2)$ using (98)]

$$= K A a E_{xz} (0) + A K A a E_{xz} (0) + A^2 E_{xz} (-3).$$

(101)

Repeating the same technique, $E_{xz}(-1)$ will take the following form,

$$E_{xz}(-1) = \sum_{i=0}^{\infty} A^i K C a A^{i+1} E_{xa} (0) C a^T.$$  

(102)

$E_{xz}(-1)$ can be evaluated numerically by taking a large number of terms for the summation (102), until the rest of the terms become negligible. $E_{x_p x_p}(0) = E \left[ x_{k-1|k-1} x_{k-1|k-1}^T \right]$ is evaluated using (97) as

$$E_{x_p x_p}(0) = K E \left[ z_k z_k^T \right] + A E \left[ x_{k-2|k-2} z_{k-1}^T \right]$$

$$+ (A E \left[ x_{k-2|k-2} z_{k-1}^T \right] + A E \left[ x_{k-2|k-2}^T z_{k-1} \right]) A^T$$

$$+ (I_n - KC) B E \left[ e_{k-2} e_{k-2}^T \right] B^T (I_n - KC)^T.$$  

(103)

Therefore, $E_{x_p x_p}(0)$ is the solution to the following Lyapunov equation,

$$A E_{x_p x_p}(0) A^T - E_{x_p x_p}(0) + K E_{z z}(0) K^T$$

$$+ A E_{x_z}(1) K^T + (A E_{x_z}(1) K)^T + (I_n - KC) B E_{z z}(0) (I_n - KC)^T = 0,$$

(104)

$E_{x_p x_p}(0)$ is divided into two parts, $\Sigma_{x_p x_p}$ and $\Sigma_{x_p x_e}$, which are independent of the watermarking signal and the false observations, respectively. $\Sigma_{x_p x_p}$ and $\Sigma_{x_p x_e}$ are the solution to the following Lyapunov equations,

$$A \Sigma_{x_p x_p} A^T - \Sigma_{x_p x_p} + K E_{z z}(0) K^T + A E_{x_z}(1) K^T$$

$$+ (A E_{x_z}(1) K^T)^T = 0,$$

$$A \Sigma_{x_p x_e} A^T - \Sigma_{x_p x_e} + (I_n - KC) B E_{z z}(0) (I_n - KC)^T = 0,$$

and $E_{x_p x_p}(0) = \Sigma_{x_p x_p} + \Sigma_{x_p x_e}$. (105)

Using (99) and (105), we can rewrite the expression for $\Sigma_{\hat{y}}$ as,

$$\Sigma_{\hat{y}} = E_{\hat{y}} (0) - C (A + BL) E_{\hat{y}} (-1)$$

$$- [C (A + BL) E_{\hat{y}} (-1)]^T + CB \Sigma_{\hat{e}} B^T C^T$$

$$+ C (A + BL) \Sigma_{\hat{e}} B (A + BL)^T C^T$$

$$+ C (A + BL) \Sigma_{\hat{e}} C (A + BL)^T C^T.$$  

(106)

**APPENDIX D**

**PROOF OF COROLLARY 1.2**

We can simplify $E_{xz} (-1)$ with the assumption that both $A$ and $A a$ are diagonalizable. If $A a$ and $A a$ are diagonalizable, then the $i$-th element of the expression for $E_{xz} (-1)$, i.e., $A a K A a^i E_{xz} (0)$, will take the following form,

$$U_a \sum_{i=0}^{\infty} \sum_{j=1}^{\infty} U_a^i U_a^{-1} A a E_{xz} (0),$$

$U_a$ replaced by eigenvalue decompositions, (47) and (48)

$$= U_a \sum_{i=0}^{\infty} \sum_{j=1}^{\infty} U_a^i U_a^{-1} A a E_{xz} (0),$$

[where $T = U_a^{-1} T U_a$, $T a$ is defined as

$$T a \triangleq \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \Sigma_{j,k} T a.$$

(109)

The $j,k$-element of the matrix $T a$ will be as follows

$$[T a]_{j,k} = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \Sigma_{j,k} T a,$$

where $[\cdot]_{j,k}$ denotes the $j$-th row and $k$-th column element of a matrix, $\lambda_{A a,j}$ and $\lambda_{a,k}$ are the $j$-th and $k$-th diagonal element of the diagonal matrices $\Sigma_{A a}$ and $\Sigma_{a}$ respectively. We assume $A a$ and $A a$ to be strictly stable, therefore, $|\lambda_{A a,j}| < 1$ and $|\lambda_{a,k}| < 1$. $|\cdot|$ denotes the absolute value of a scalar. Using (109), we can write

$$E_{xz} (-1) = U a T a U a^{-1} A a E_{xz} (0).$$  

(110)
APPENDIX E
Proof of Theorem 2

This section provides the proof of the Theorem 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 [39]. Using (36), (37) and (87), and considering that $e_k$ and $w_{a_k}$ are uncorrelated with $z_k$ and $\hat{x}_{k|k}$, and also with each other, we can write,

$$\Sigma_k = Q_a + E \left[ \mu_{g_k|\gamma_1}^{(1)}, (\epsilon_k^{1})^{T} \mu_{g_k|\gamma_1}^{(1)}, (\epsilon_k^{1})^{T} \right].$$

(111)

The expected KLD $E \left[ D \left( f_{g_k}^{(1)}, f_{\gamma_1}^{(1)} \right) \right]$ under the optimal CUSUM test is derived as follows using (39), see (112).

Similarly, the KLD $D \left( f_{\gamma_1}^{(1)}, f_{g_k}^{(1)} \right)$ under the sub-optimal CUSUM test will take the following form [39],

$$\frac{1}{2} \left( \log \frac{\Sigma_{g_k}}{\Sigma_{\gamma_1}} - p - m - tr \left( \Sigma_{\gamma_1}^{-1} \Sigma_{g_k} \right) \right).$$

(113)

The term log $|\Sigma_{\gamma_1}|$ is evaluated as follows,

$$|\Sigma_{\gamma_1}| = |\Sigma_e| |\Sigma_{\gamma_1}| \text{, [using (40)]}$$

(114)

$$|\Sigma_{g_k}| = |\Sigma_e| |\Sigma_e - CB \Sigma_e B^T C^T|, \text{ [using (41)].}$$

(115)

Therefore, log $|\Sigma_{\gamma_1}|$ = log $|\Sigma_e - CB \Sigma_e B^T C^T|$.

(116)

The term $tr \left( \Sigma_{\gamma_1}^{-1} \Sigma_{g_k} \right)$ is evaluated using (40) and (41) as,

$tr \left( \Sigma_{g_k}^{-1} \Sigma_{\gamma_1} \right) = tr \left( \Sigma_{g_k}^{-1} \Sigma_{\gamma_1} + \Sigma_{g_k}^{-1} \Sigma_{g_k} \right) = tr \left( \Sigma_{\gamma_1} \Sigma_{\gamma_1} \right) + p$

(117)

Applying (116) and (117) in (113), we get the final expression of the KLD $D \left( f_{\gamma_1}^{(1)}, f_{g_k}^{(1)} \right)$ under the sup-optimal CUSUM test as

$$\frac{1}{2} \left( tr \left( \Sigma_{\gamma_1}^{-1} \Sigma_{g_k} \right) - m - log \left| \Sigma_e - CB \Sigma_e B^T C^T \right| \right).$$

(118)

APPENDIX F
Proof of Lemma 2

This section provides the derivation of the expression of $\sigma^2_\gamma$ for the MISO system. The model parameters of the fake measurement generation system [12] for the MISO system will be as follows.

$$A_a = \rho, \quad Q_a = (1 - \rho^2) \sigma^2_z, \quad \text{and} \quad E_{zz}(0) = \sigma^2_z.$$  

(119)

To evaluate $\sigma^2_\gamma$, we derive the expression for $E_{zz}(-1)$ for a MISO system using [43] as

$$E_{zz}(-1) = \sum_{i=0}^{\infty} A^i KA_a^{i+1} E_{zz}(0) = \sum_{i=0}^{\infty} A^i K \rho^{i+1} \sigma^2_z \left[ E_{zz}(0) = \sigma^2_z, \quad A_a = \rho \right] = \left[ I_n - \rho A \right]^{-1} K \rho \sigma^2_z, \quad [A \text{ is strictly stable, } \rho < 1].$$  

(120)

$\sigma^2_\gamma$ will be as follows,

$$\sigma^2_\gamma = \sigma^2_z - 2C (A + BL) E_{zz}(-1) + CB \Sigma_z B^T C^T$$

$$+ C (A + BL) \Sigma_{zfz} (A + BL)^T C^T$$

$$+ C (A + BL) \Sigma_{ze} (A + BL)^T C^T \text{ [using (42)],}$$

(121)

where $\Sigma_{zfz}$ and $\Sigma_{ze}$ are derived from (44) and (45) respectively as follows.

$$\Sigma_{zfz} = \Sigma_{zfz}^2 \sigma^2_z$$

(122)

where $\Sigma_{zfz}$ is the solution to the following Lyapunov equation,

$$A \Sigma_{zfz} A^T - \Sigma_{zfz} + K K^T + A [I_n - \rho A]^{-1} K K^T \rho$$

$$+ \left[ A [I_n - \rho A]^{-1} K K^T \rho \right]^T = 0.$$  

(123)

$\Sigma_{ze}$ is the solution to the following Lyapunov equation,

$$A \Sigma_{ze} A^T - \Sigma_{ze} + (I_n - KC) B \Sigma_e B^T (I_n - KC)^T = 0.$$  

(124)

Using (120) and (122), the expression for $\sigma^2_\gamma$ [12] can be rearranged as follows.

$$\sigma^2_\gamma = \left( 1 - 2C (A + BL) (I_n - \rho A)^{-1} K \rho \right)$$

$$+ C (A + BL) \Sigma_{zfz} (A + BL)^T C^T \sigma^2_z$$

$$+ \left( C (A + BL) \Sigma_{ze} (A + BL)^T C^T + CB \Sigma_e B^T C^T \right)$$

$$= M_1 \sigma^2_z + M_1$$

(125)

The scalar quantity $M_1$ can be rearranged as follows.

$$M_1 = \left( \sum_{l=0}^{\infty} C (A + BL) A^t (I_n - KC)^T \sum_{l=0}^{\infty} B^T (I_n - KC)^T [A^t]^t (A + BL)^T C^T (A + BL) \right)$$

$$A^t (I_n - KC)^T B \Sigma_e B^T C^T - \text{tr} \left( M_1 \Sigma_e \right),$$

(126)

where $M_1 = B^T (I_n - KC)^T \Sigma_{zfz} (I_n - KC)^T B + B^T C^T CB.$

(127)

$\Sigma_{zfz}$ is the solution to the following Lyapunov equation,

$$A^t \Sigma_{zfz} A - \Sigma_{zfz} + (A + BL)^T C^T C (A + BL) = 0.$$  

(128)

Finally, we can write $\sigma^2_\gamma$ as

$$\sigma^2_\gamma = M_1 \sigma^2_z + \text{tr} (M_1 \Sigma_e).$$

(129)

APPENDIX G
Proof of Theorem 4

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

$$\Sigma_e = V_e A_e V_e^T$$  

(130)

where $V_e$ and $A_e$ are the eigenvector matrix and the diagonal eigenvalue matrix. In this section, we will prove that KLD is
We have replaced all the B matrices by $B_e$ where $B_e = BV_e$ and $\Sigma_e$ by $\Lambda_e$ to keep the structure of the KLD and $\sigma^2_\gamma$ expressions as (62) and (64) respectively.

$$f(\Lambda_e) = \frac{1}{2} \left( M_e \sigma^2_\gamma + \sum_{i=1}^n [M_{e,\lambda}]_{ii} \lambda_e,i \right)$$

$$- \frac{1}{2} \log \left( \frac{M_e \sigma^2_\gamma + \sum_{i=1}^n [M_{em}]_{ii} \lambda_e,i}{\sigma^2_\gamma} \right)$$

(139)

where $M_{em} = B^T_e (I_n - KC)^T \Sigma_{\varepsilon,\rho} (I_n - KC) B_e$, and $M_{e,\lambda} = B^T_e (I_n - KC)^T \Sigma_{\varepsilon,\rho} (I_n - KC) B_e + B^T_e C^T CB_e$.

(140)

The $\Sigma_{\varepsilon,\rho}$ is the same as in (128). The first derivative of the cost function with respect to the $j$-th eigenvalue $\lambda_{e,j}$ is as follows,

$$\frac{\partial}{\partial \lambda_{e,j}} f(\Lambda_e) = \frac{1}{2} \left[ M_{e,\lambda} \right]_{jj}$$

(141)

The second derivative of the cost function is as follows,

$$\frac{\partial^2}{\partial \lambda_{e,i} \partial \lambda_{e,j}} f(\Lambda_e) = \frac{1}{2} \left[ M_{em} \right]_{ii} \left[ M_{em} \right]_{jj} t^2_{f,ij}$$

(142)

where $\frac{\partial}{\partial \lambda_{e,i} \partial \lambda_{e,j}} f(\Lambda_e)$ is the $ij$-th element of the Hessian matrix $H_s = \nabla^2 f(\Lambda_e)$. From (142), it is clear that each column of $H_s$ 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 $H_s$ are zero. Also, the diagonal elements of $H_s$ are non-zero. So, we can conclude that KLD is convex in $\Lambda_e$.

Since the cost function under both the tests are convex, the optimum $\Lambda_e$, which maximizes the expected KLD or the KLD, will be on one of the vertices of the feasible region provided by (132) and (133). That is possible when the optimum $\Lambda_e$ 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**

**OPTIMIZATION ALGORITHM**

The Lagrangian and it’s first and second derivatives for the MISO system are given as follows. We multiply the cost function by -1 to convert the optimization problem into a minimization one.
For the optimal CUSUM test, using (77) and (57) the Lagrangian can be written in the following form
\[ L(v_\lambda, \mu) = -v_\lambda^T H_{KL} v_\lambda + \mu \left( v_\lambda^T H v_\lambda - J \right). \] (143)

The first derivatives of \( L(v_\lambda, \mu) \) with respect to \( v_\lambda \) and \( \mu \) are
\[ \nabla v_\lambda L(.) = C_c v_\lambda, \quad \text{and} \quad \frac{\partial}{\partial \mu} L(.) = v_\lambda^T H v_\lambda - J \] (144)
where \( C_c = -2H_{KL} + 2\mu H \). (146)
The Hessian matrix of \( L(.) \) with respect to \( v_\lambda \) is as follows,
\[ H_s = \nabla^2 v_\lambda L(.) = C_c^T. \] (147)

For the sub-optimal CUSUM test, we form the Lagrangian using (62), (66), and (57) for the KLD and \( \Delta LQG \) respectively as follows
\[ L(v_\lambda, \mu) = -\frac{1}{2} \left( M_x \sigma_z^2 + v_\lambda^T M_{ex} v_\lambda + v_\lambda^T B^T C^T C B v_\lambda \right) \] (148)
\[ + \frac{1}{2} \log \left( M_x \sigma_z^2 + v_\lambda^T M_{ex} v_\lambda \right) - \frac{1}{2} \log \left( \sigma_z^2 \right) \]
\[ + \mu \left( v_\lambda^T H v_\lambda - J \right), \]
where \( M_{ex} \) is the first part of the right hand side of (66), i.e., \( M_{ex} = B^T (I_n - K C)^T \Sigma_{ex} \) \((I_n - K C) B \). (150)

The Hessian matrix of \( L(.) \) with respect to \( v_\lambda \) is as follows
\[ H_s = \nabla^2 v_\lambda L(.) = C_{ca}^T + 2M_{ex} v_\lambda v_\lambda^T C_{cb}. \] (151)

A primal-dual approach to find the optimal \( \Sigma_e \) is provided in Algorithm 1. The step sizes \( (s_k, K_{\mu,k}) \) can be derived at every step using the backtracking algorithm [40] which ensures the convergence to some local optima since the Hessian matrices under both the tests are indefinite matrices.

**Algorithm 1** To find optimum \( \Sigma_e \)

1. Initialize: \( s_0, K_{\mu,0}, \text{max}_{\text{iteration}}, \text{and} \mu = 0. \)
2. for \( k = 1 : \text{max}_{\text{iteration}} \) do
   
   Find the best solution \( v_{\text{temp}}^* \) for the set of equations, \( \nabla v_\lambda L(.) = 0 \) and \( \frac{\partial}{\partial \mu} L(.) = 0. \)
   
   if \( v_{\text{temp}}^T H_{\text{temp}} v_{\text{temp}}^* = 0 \) then
     \[ \mu \leftarrow \mu + s_k \frac{\partial}{\partial \mu} L(.) \]
   else
     \[ \mu \leftarrow K_{\mu,k} \left( -\frac{\partial}{\partial \mu} L(.) \right) \]
   end if
3. end if
4. end for

**System-B parameters:**
\[ A = \begin{bmatrix} 0.968 & 0 & 0.082 & 0 \\ 0 & 0.978 & 0 & 0.064 \\ 0 & 0 & 0.917 & 0 \\ 0 & 0 & 0 & 0.935 \end{bmatrix}, \quad B = \begin{bmatrix} 0.164 & 0.004 \\ 0.002 & 0.124 \\ 0 & 0.092 \end{bmatrix}, \quad C = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \end{bmatrix}, \quad R = \text{diag} [0.5, 0.5] \]
\[ Q = \text{diag} [0.25, 0.25, 0.25, 0.25], \quad U = \text{diag} [2, 2], \quad W = \text{diag} [5, 5, 5], \quad Q_a = \text{diag} [0.4, 0.2, 0.2, 0.7] \]

**REFERENCES**

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

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

[3] M. Abrams and J. Weiss, “Malicious Control System Cyber Security Attack Case Study – Maroochy Water Services, Australia,” MITRE Corp USA, vol. 253, no. August, pp. 73–82, 2008.

[4] A. Cardenas, S. Amin, B. Sinopoli, A. Giani, A. Perrig, and S. Sastry, “Challenges for Securing Cyber Physical Systems,” 2009.

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

[6] S. Salimi, S. Dey, and A. Ahlen, “Sequential Detection of Deception Attacks in Networked Control Systems with Watermarking,” pp. 883–890, 2019.

[7] H. Fawzi, P. Tabuada, and S. Diggavi, “Secure estimation and control for cyber-physical systems under adversarial attacks,” IEEE Trans. Automat. Contr., vol. 59, no. 6, pp. 1454–1467, 2014.

[8] D. Du, X. Li, W. Li, R. Chen, M. Fei, and L. Wu, “ADMM-Based Distributed State Estimation of Smart Grid under Data Deception and Denial of Service Attacks,” IEEE Trans. Syst. Man, Cybern. Syst., vol. 49, no. 8, pp. 1698–1711, 2019.

[9] N. Forti, G. Battistelli, L. Chisci, S. Li, B. Wang, and B. Sinopoli, “Distributed Joint Attack Detection and Secure State Estimation,” IEEE Trans. Signal Inf. Process. over Networks, vol. 4, no. 1, pp. 96–110, mar 2018.

[10] G. Park, C. Lee, H. Shim, Y. Eun, and K. H. Johansson, “Stealthy Adversaries against Uncertain Cyber-Physical Systems: Threat of Robust Zero-Dynamics Attack,” IEEE Trans. Automat. Contr., vol. 64, no. 12, pp. 4907–4919, dec 2019.
[11] Y. Chen, S. Kar, and J. M. Moura, “Cyber-Physical Attacks with Control Objectives,” IEEE Trans. Automat. Contr., vol. 63, no. 5, pp. 1418–1425, may 2018.
[12] F. Pasqualetti, F. Dorfler, and F. Bullo, “Attack detection and identification in cyber-physical systems,” IEEE Trans. Automat. Contr., vol. 58, no. 11, pp. 2715–2729, nov 2013.
[13] E. Moukasnejad, Y. Yang, Q. L. Han, and L. Vlacic, “A novel cyber attack detection method in networked control systems,” IEEE Trans. Cybern., vol. 48, no. 11, pp. 3254–3264, nov 2018.
[14] X. Ge, Q. L. Han, M. Zhong, and X. M. Zhang, “Distributed Krein space-based attack detection over sensor networks under deception attacks,” Automatica, vol. 109, pp. 108557, sep 2019.
[15] Y. Mo and B. Sinopoli, “Secure control against replay attacks,” 2009 47th Annu. Allert. Conf. Commun. Control. Comput. Allert, pp. 911–918, sep 2009.
[16] Y. Mo, R. Chabukswar, and B. Sinopoli, “Detecting integrity attacks on SCADA systems,” IEEE Trans. Control Syst. Technol., vol. 22, no. 4, pp. 1396–1407, 2014.
[17] W. H. Ko, B. Satchidanandan, and P. R. Kumar, “Dynamic watermarking-based defense of transportation cyber-physical systems,” ACM Trans. Cyber-Physical Syst., vol. 4, no. 1, 2019.
[18] C. Fang, Y. Qi, P. Cheng, and W. X. Zheng, “Optimal periodic watermarking schedule for replay attack detection in cyber–physical systems,” Automatica, vol. 112, p. 108698, 2020.
[19] B. Satchidanandan and P. R. Kumar, “On the Design of Security-Guaranteeing Dynamic Watermarks,” IEEE Control Syst. Lett., vol. 4, no. 2, pp. 307–312, 2020.
[20] P. Hespanhol, M. Porter, R. Vasudevan, and A. Aswani, “Statistical Watermarking for Networked Control Systems,” Proc. Am. Control Conf., vol. 2018-June, pp. 5467–5472, 2018.
[21] S. Weerakkody, O. Ozel, and B. Sinopoli, “A Bernoulli-Gaussian physical watermark for detecting integrity attacks in control systems,” 55th Annu. Allert. Conf. Commun. Control. Comput. Allert, vol. 2018-Janua, no. Iid, pp. 966–973, 2018.
[22] P. Pradhan and P. Venkitasubramaniam, “Stealthy Attacks in Dynamical Systems: Tradeoffs between Utility and Detectability with Application in Anonymous Systems,” IEEE Trans. Inf. Forensics Secur., vol. 12, no. 4, pp. 779–792, 2017.
[23] A. N. Shiryaev, “On optimum methods in quickest detection problems,” Theory Probab. Its Appl., vol. 8, no. 3, pp. 456–465, may 2000.
[24] A. N. Shiryaev and S. Amin, “Secure control: Towards survivable cyber-physical systems,” Proc. - Int. Conf. Distrib. Comput. Commun., vol. 2018-June, pp. 441–475, 2018.
[25] A. Tartakovsky, “On Asymptotic Optimality in Sequential Change-Point Detection: Non-iid Case,” IEEE Trans. Inf. Theory, vol. 63, no. 6, pp. 3433–3450, 2017.
[26] T. L. Lai, “Information bounds and quick detection of parameter changes in stochastic systems,” IEEE Trans. Inf. Theory, vol. 44, no. 7, pp. 7917–7929, 1998.
[27] A. Tartakovsky, I. Nikiforov, and M. Basseville, Sequential analysis: Hypothesis testing and change-point detection, 2014.
[28] V. Girardin, V. Konev, and S. Pergamenchtchikov, “Kullback-Leibler Approach to CUSUM Quickest Detection Rule for Markovian Time Series,” Seq. Anal., vol. 37, no. 3, pp. 322–341, 2018.
[29] A. A. C árdenas, S. Amin, and S. Sastry, “Secure control: Towards survivable cyber-physical systems,” Proc. - Int. Conf. Distrib. Comput. Syst., pp. 495–500, 2008.
[30] S. Salimi, S. Dey, and A. Ahlen, “Sequential detection of deception attacks in networked control systems with watermarking,” 2019 18th Eur. Control Conf. ECC 2019, pp. 883–890, 2019.
[31] A. Naha, A. Teixeira, A. Ahlen, and S. Dey, “Sequential detection of replay attacks,” arXiv preprint, arXiv:2012.10748, 2020. [Online]. Available: arXiv:2012.10748
[32] D. I. Urbina, J. Giraldo, A. A. Cardenas, N. O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, and H. Sandberg, “Limiting the impact of stealthy attacks on Industrial Control Systems,” in Proc. ACM Conf. Comput. Commun. Secur., 2016, pp. 1092–1105.
[33] J. Giraldo and A. A. Cardenas, “A new metric to compare anomaly detection algorithms in cyber-physical systems,” in Proc. 6th Annu. Symp. Hot Top. Sci. Secur., 2019, pp. 1–2.
[34] P. T. Boggs and J. W. Tolle, “Sequential Quadratic Programming,” Acta Numer., vol. 4, no. 1995, pp. 1–51, 1995.
[35] A. Forsgren, P. E. Gill, and M. H. Wright, “Interior methods for nonlinear optimization,” SIAM Rev., vol. 44, no. 4, pp. 525–597, 2002.
[36] K. H. Johansson and J. L. R. Nunes, “The Quadruple-Tank Process: A Multivariable Laboratory Process with an Adjustable Zero,” Proc. Am. Control Conf., vol. 8, no. 3, pp. 456–465, may 2000.