Quickest Detection of Deception Attacks on Cyber-Physical Systems with a Parsimonious Watermarking Policy

Arunava Naha\textsuperscript{a,∗}, André Teixeira\textsuperscript{b}, Anders Ahlén\textsuperscript{a} and Subhrakanti Dey\textsuperscript{a}
\textsuperscript{a}Electrical Engineering, Uppsala University, Uppsala, Sweden
\textsuperscript{b}Department of Information Technology, Uppsala University, Uppsala, Sweden

Abstract

Adding a physical watermarking signal to the control input of a networked control system increases the detection probability of data deception attacks at the expense of increased control cost. This paper proposes a parsimonious policy to limit the average number of watermarking events when the attack is not present, which in turn reduces the control cost. We model the system as a stochastic optimal control problem and apply dynamic programming to minimize the average detection delay (ADD) for fixed upper bounds on false alarm rate (FAR) and an average number of watermarking events (ANW) before the attack. Under practical circumstances, the optimal solution results in a two threshold policy on the posterior probability of attack, derived from the Shiryaev statistics for sequential change detection and assuming the change point is a random variable. We derive asymptotically approximate analytical expressions of ADD and FAR, applying the non-linear renewal theory for non-independent and identically distributed data. The derived expressions reveal that ADD reduces with the increase in the Kullback-Leibler divergence (KLD) between the post- and pre-attack distributions of the test statistics. Therefore, we further design the optimal watermarking that maximizes the KLD for a fixed increase in the control cost. The relationship between the ANW and the increase in control cost is also derived. Simulation studies are performed to illustrate and validate the theoretical results.

Key words: attack detection, cyber-physical system, deception attack, Kullback–Leibler divergence, linear quadratic Gaussian control, networked control system, physical watermarking, sequential change detection in Bayesian setting, Shiryaev statistics.

I. Introduction

Nowadays, cyber-physical systems (CPS) with embedded software, processors, sensor network, and other physical components are getting deployed for advanced healthcare, smart buildings, smart manufacturing units, intelligent transport systems, defence purposes, smart grids, etc. [1]. CPS integrate cyber and physical components by exchanging data over the wireless network and provide autonomy, reliability, accuracy, and real-time control without human involvement [1]. [2]. Along with their numerous advantages, there is also a growing concern regarding the safety and security of CPS. Due to the use of commodity software and off-the-shelf networking components, unattended operations, and a few other reasons CPS are vulnerable to adversarial attacks on the cyber or/and physical layer [3]. Cryptography, firewalls, user authentications, digital watermarking, etc. are already in place to protect CPS from cyber attacks. However, such protection mechanisms may not be adequate against physical attacks. For example, during the Stuxnet attack [4], attackers issued harmful exogenous control inputs to increase the pressure of the centrifuges beyond the safety limit at a uranium enrichment plant in Iran. To remain stealthy during the attack, attackers also replaced the true observation from the system with previously recorded data. There are a few other examples, such as the attack on a sewage system in Australia [5], the attack on the Davis-Besse nuclear power plant in Ohio, USA [6], etc., where cyber protection schemes failed to prevent or detect the attacks. Attacks on CPS may cause considerable monetary loss and pose threats to human safety [1].

Attack strategies for the physical layer of CPS can be broadly classified into two groups, data deception attacks and denial of service (DoS) attacks. In data deception attacks, the adversary feeds the system with false data [1], [3]. Replay attacks are one kind of data deception attack, where the attacker replaces the true observations with previously recorded data to remain stealthy [4]. In DoS attacks, the attacker’s objective is to disrupt the availability of data. The attacker may achieve that by overpowering the wireless network [7]. In an attack scenario, the attacker’s objective is to remain stealthy as long as possible and cause maximum damage to the system. The inherent noise and uncertainties in CPS assist the attacker in achieving such an objective. The role of a control system engineer is to detect the attack as soon as possible to minimize the damage. In this paper, we have studied data deception attacks on networked control systems (NCS), where the attacker replaces the true observation with fake data.

A. Related Work

Researchers are working on different challenges to secure CPS against attacks on the physical layer, such as the study of different attack strategies [8], [9], attack resilient state estimation [10], [12], attack detection strategies [1], [3], [13]–[19], etc. Detection strategies for the attacks on the physical layer of CPS can be broadly divided into two groups, passive and active.
schemes. Under the passive attack detection scheme, the innovation signal from the state estimator or the observation signal is subjected to various statistical tests [13]–[15]. However, as studied in the literature, passive detection schemes generally have an unsatisfactory probability of detection in the presence of noise and uncertainties [20]. On the other hand, under the active attack detection scheme, physical watermarking signals are added to the control inputs, and various statistical tests are used to check the authenticity of the received observations. The physical watermarking scheme was first introduced in [20] to detect replay attacks by adding an iid watermarking signal to the control input and performing a $\chi^2$ test using the innovation signal from the state estimator. The method in [20] is improved by designing an optimal watermarking signal in [17]. Instead of an iid watermarking scheme, the watermarking signal generated from a hidden Markov Model (HMM) is studied in [3]. A sequential attack detection scheme using the CUSUM statistics evaluated from the joint distribution of the added watermarking and the innovation signal is studied in [21]. Besides the innovation signal, the observation signal is also used to generate residue signals for the attack detections [1]. In [1], [3], [21], watermarking signals are added to the control inputs for all the time instants till the point of attack detection. The addition of physical watermarking to the control input increases the probability of attack detection at the expense of increased control cost [1]. The relation between the increase in the linear quadratic Gaussian control cost, $\Delta LQG$, and the watermarking signal variance is studied in [3]. Since the attack is a less frequent event, adding the watermarking signal during the normal operation for a long time can increase the total control cost significantly [18] and unnecessarily. In the current paper, we have studied an evidence-based watermarking policy to reduce the increase of control cost before an attack, and, at the same time, achieve satisfactory detection performance.

Researchers are exploring diverse approaches to reduce the increase in the control cost due to the added watermarking and maintain satisfactory detection performance. In one approach, the authors have added the watermarking periodically to the control inputs and kept a balance between the improvement in the control cost and the increase in the detection delay [18]. Another approach is to add watermarking directly to the observations [22]–[24]. In this approach, the authenticity of the observations is first examined at the receiving end, and then the watermarking signal is filtered out before using the observations in the controller. Since the watermarking signal is filtered out, the control cost does not increase. Different kinds of watermarking signals are used in this context, such as sinusoidal [25], time-varying sinusoidal [26], random noise [24], multiplicative to the observations [22], etc. However, these methods may fail in the scenario, where the attacker hijacks the sensor node and feeds the fake data before the addition of the watermarking. In general, the physical watermarking-based methods targeting to reduce the increased control cost or more traditional always present watermarking-based methods use batch processing of data, i.e., innovation signal or observation signal. Therefore, those methods do not address the problem of the quickest attack detection. However, we know that early detection of attacks is of paramount importance for CPS to reduce the amount of damage.

Therefore, in this paper, we studied the problem of the quickest detection of attacks which uses watermarking parsimoniously to reduce the loss in control performance prior to an attack. The literature on the quickest detection of a change point by sequential analysis of data dates back several decades. A brief review on the quickest change detection techniques is provided in the following paragraph.

The quickest change detection methods can be classified into two broad groups depending upon the assumption of the model of the change point [26]. In one approach, which is also called the minimax approach, the change point is modelled as deterministic but unknown. The cumulative sum (CUSUM) technique is one of such minimax approaches, which was first introduced by Lorden [27]. In the other approach, the Bayesian approach, the change point is modelled as a random variable (RV) with some prior distribution. The Bayesian change point detection technique was first introduced by Shiryaev [28]. The original Shiryaev rule was proposed for the data with different iid distributions, before and after the change point. Finding an optimal detection rule for the general non-iid data is difficult [29]. In [30], an optimal detection rule is developed for homogeneous finite-state Markov chains. A slightly different approach is followed in [31], [32], where the authors proved that the Shiryaev rule, with minor modifications, is an asymptotically optimal quickest change detection rule under the conditions given in (3), (4), (21) and (23) of [32]. That means the Shiryaev procedure minimizes the average detection delay (ADD) for a fixed upper threshold on the false alarm rate (FAR) for the non-iid data provided that the threshold $\rightarrow \infty$ and a few other conditions are satisfied. The condition (3) of [32] for the optimality is that the prior distribution of the change point must satisfy

\[
\lim_{k \rightarrow \infty} \frac{\log P \{ \Gamma \geq k + 1 \}}{k} = -c, \quad c \geq 0,
\]

where $\Gamma$ is the change point. That means the exponential rate of convergence of the prior distribution must be $c \geq 0$, where $c > 0$ indicates the prior distribution has an exponential right tail, and $c = 0$ indicates the prior distribution is heavy-tailed [29]. An attacker will always try to remain stealthy for a long time because the longer time the attacker remains undetected, the more damage can be caused [18]. On the other hand, the defender should design a detection mechanism that will detect the attack as soon as possible with an acceptable FAR to reduce the amount of damage. Therefore, we have used the Bayesian approach in this paper, which minimizes the ADD, whereas the other method, the minimax approach, only minimizes the worst-case ADD (computed over all possible attack start points) [26]. Additionally, our work in this paper is inspired by two other prior works [33], [34]. The quickest intrusion detection problem is studied in [33], where only a minimal set of sensors from a sensor network is kept active at a particular time instant. The problem of quickest change detection is also studied in [34] with upper bounds on the average number of sensor data used before the change point and the FAR. In both the problem formulations, the underlying data was assumed to be iid, which is not the case for the system under study in this paper. However, similar to several other works on change-point detections [33], [34], we have also assumed the distribution of the change point, i.e., the
attack start point, to be a geometric distribution with parameter \( \rho \), which satisfies the condition given in [1].

### B. Contributions

In our previous work [21], we studied in detail that the worst-case ADD decreases with the increase in \( \Delta LQG \), which denotes the increases in the LQG control cost due to the addition of watermarking for the always-present watermarking scheme. Additionally, \( \Delta LQG \) is proportional to the watermarking signal power. In other words, an attack can be detected early with higher watermarking signal power, i.e., at the expense of increased control cost. Therefore, in this paper, we propose a method to reduce the average number of watermarking (ANW) events used before the attack start point, which reduces the average watermarking signal power and subsequently \( \Delta LQG \).

We formulate the task at hand as a stochastic optimal control problem to minimize the ADD for fixed upper bounds on FAR and ANW and apply dynamic programming to solve it. Similar to any other detection technique, there is always a trade-off between ADD and FAR [31]. We have studied the structure of the dynamic programming solution, i.e., the solution of the Bellman equation, and found that the optimal policy is a two threshold policy with thresholds \( T h^a \) and \( T h^d \), \( T h^d \geq T h^a \), on the posterior probability of attack \( p_k \) under practical circumstances. In other words, if \( p_k \geq T h^a \), then we add watermarking to the \( (k + 1) \)-th control input. On the other hand, if \( p_k \geq T h^d \), we decide that the attack is present in the system and terminate the process. Our study shows that \( T h^a \) primarily controls the ANW, which in turn controls the \( \Delta LQG \) value, and \( T h^d \) primarily controls the ADD and FAR.

Asymptotically approximate analytical expressions of ADD and FAR are derived by applying non-linear renewal theory for non-iid data. The derived expression of ADD indicates that the ADD reduces with the increase of the Kullback-Leibler divergence (KLD) between the post- and pre-attack distributions of the test data. Additionally, the derived analytical expression of KLD for our problem formulation provides the relationship between the KLD and the watermarking signal variance. Therefore, we use this relationship to derive the optimal watermarking signal variance, which maximizes the KLD for a given upper bound on \( \Delta LQG \). We have also obtained an expression of \( \Delta LQG \) for a given ANW. We have reported a preliminary simulation study on this problem previously for a single-input single-output (SISO) system in [35]. In the current paper, we have performed a more in-depth theoretical analysis of the problem for general multi-input and multi-output (MIMO) system models. Our main contributions are as follows.

1. To the best of our knowledge, this is the first time the Bayesian approach is applied for the quickest detection of data deception attacks on NCS with a parsimonious watermarking policy to reduce the control cost.
2. We have derived asymptotically approximate analytical expressions of ADD, FAR and \( \Delta LQG \) that facilitate the optimal design of the watermarking process.
3. We have optimised the watermarking signal variance to maximise KLD, which improves ADD for a fixed upper bound on the \( \Delta LQG \).

The paper is organized as follows. Section II discusses the system model and the attack strategy considered in this paper. The defence mechanism is explained in Section III. Section IV provides the analytical expressions of ADD, FAR and the relationship between the ANW and \( \Delta LQG \). It also explains the optimization framework for the watermarking signal variance. Section V presents and discusses the simulation results. Section VI concludes the paper.

### C. 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.

#### II. System Model

The system model during normal operations and the model with the data deception attack are discussed in this section.

#### A. System model during normal operation

A schematic diagram of a standard NCS during the normal operation is shown in Fig. 1. We assume a linear time-invariant system model and the attack strategy considered in this paper.

![Schematic diagram of the system during normal operation](Image)

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

MIMO plant with the following state update and measurement equations,

\[
x_{k+1} = \mathbf{A}x_k + \mathbf{B}u_k + \mathbf{w}_k,
\]

\[
y_k = \mathbf{C}x_k + \mathbf{v}_k,
\]

where \( x_k \in \mathbb{R}^n \) and \( y_k \in \mathbb{R}^m \) are the state and measurement vectors, respectively. \( u_k \in \mathbb{R}^p \) is the control input vector. The process and observation noise vectors are denoted as \( \mathbf{w}_k \in \mathbb{R}^n \) and \( \mathbf{v}_k \in \mathbb{R}^m \), respectively.
\( \mathbb{R}^n \sim \mathcal{N}(0, Q) \) and \( y_k \in \mathbb{R}^m \sim \mathcal{N}(0, R) \), respectively, with \( Q > 0 \) and \( R > 0 \). Here, \( A \in \mathbb{R}^{n \times n} \), \( B \in \mathbb{R}^{n \times p} \), \( Q \in \mathbb{R}^{n \times n} \), \( C \in \mathbb{R}^{m \times n} \), and \( R \in \mathbb{R}^{m \times m} \). Process and observation noises are assumed to be iid and uncorrelated to each other and with the initial state vector. We also assume that the system has been operational for a very long time and is currently in steady-state.

The states of the system are estimated using the Kalman estimator. The sensor measurements are available to a remote estimator/controller, possibly over a wireless link, which may be vulnerable to malicious data deception attacks. In the absence of an attack, the time update and measurement update equations (4) and (5), respectively. \( E \) are the Kalman predicted and filtered states, respectively. \( E[\cdot] \) denotes the expectation operator, and \( Z_k \triangleq \{u_0, u_1, \ldots, u_k, y_0, y_1, \ldots y_k \} \) is the set of all information up to time \( k \). The innovation signal \( \gamma_k \) and the steady state Kalman filter gain \( K \) are given as,

\[
\gamma_k = y_k - C \hat{x}_{k|k-1}, \tag{6}
\]

\[
K = PC_T (CPC_T + R)^{-1}, \tag{7}
\]

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

\[
P = APA^T + Q - APC^T (CPC^T + R)^{-1} CPA^T. \tag{8}
\]

The estimated states are fed to a state feedback controller which is assumed to be an infinite horizon linear quadratic Gaussian (LQG) controller. The optimal control input \( u_k^* \) is derived by minimizing the following cost function,

\[
J = \lim_{T \to \infty} E \left[ \frac{1}{2T+1} \sum_{k=-T}^{T} (x_k^T W x_k + u_k^T U u_k) \right]. \tag{9}
\]

Here \( W \geq 0 \) and \( U \geq 0 \) are weight matrices. The optimal LQG control input turns out to be the following linear function of the estimated states, \( u_k^* = L \hat{x}_{k|k} \), where

\[
L = -(B^T SB + U)^{-1} B^T SA. \tag{10}
\]

Here, \( S \) is the solution to the following algebraic Riccati equation,

\[
S = A^T SA + W - A^T SB (B^T SB + U)^{-1} B^T SA. \tag{11}
\]

**B. Attack Model**

We make the following assumptions regarding the capabilities and knowledge of an attacker:

1) the attacker can access the sensor nodes and replace the true observations with fake data;
2) the attacker has complete knowledge about the system and the controller, i.e., the attacker knows \( A, B, C, Q, R, \) and \( L \);
3) the attacker can not access or alter the control signal.

To launch a data deception attack, the attacker replaces the true observations \( y_k \) by the fake data \( z_k \) from \( k \geq \Gamma \). A well-studied method to achieve this is to transmit the fake observations with significantly higher power than the true measurements from the sensors. As a result, the wireless control system receiver accepts the fake measurements as legitimate while rejecting the true measurements from the sensor nodes. Such attack models are also known as sensor spoofing attacks \cite{35, 36}. The fake observation data \( z_k \) is assumed to be generated from the following linear stationary stochastic process,

\[
z_k = A_a z_{k-1} + w_{a,k-1}, \tag{11}
\]

where \( w_{a,k} \sim \mathcal{N}(0, Q_a) \) is the iid noise vector at the \( k \)-th time instant, and \( Q_a \in \mathbb{R}^{m \times m} \). A similar attack strategy is also studied in \cite{37}, where the stealthiness of the attack signal is evaluated in terms of the KLD between the distributions of the fake and true observations. The attacker’s system matrix \( A_a \) and the noise covariance matrix \( Q_a \) should be designed in such a way so that the fake data \( z_k \) mimics the statistical properties of the true measurement \( y_k \). \( A_a \) mainly models the correlations between the current and past measurements, and \( Q_a \) models the uncertainty. Designing \( A_a \) and \( Q_a \) in such a way increases the stealthiness of the attack signal. A schematic diagram of the system under the data deception attack is shown in Fig. 2.

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

**Remark 1.** In general, for linear control systems, the measurement vector \( y_k \) can be modelled as a stochastic process that is dependent on its past values with an additive i.i.d noise component, i.e., similar to \cite{38}. In other words, the attack model in \cite{38} mimics the linear stationary stochastic model of \( y_k \), which makes it challenging to detect. In addition to that, the attack model in \cite{38} can make closed-loop control systems unstable, as discussed in \cite{39}, which may cause a significant amount of damage to CPS. Since the attacker’s objectives are to cause damage to the CPS and to remain stealthy in doing so, the attack model considered in this paper \cite{38} facilitates the attacker in achieving both the objectives, thus illustrating the significance of such an attack model. Moreover, such an attack model can be used to detect replay attacks after a few modifications, as demonstrated in \cite{40}.

During the attack, i.e., for \( k \geq \Gamma \), the innovation signal will take the following form,

\[
\gamma_k = z_k - C \hat{x}_{k|k-1}, \tag{12}
\]

whereas \( \hat{x}_{k|k-1} \) and \( \hat{x}_{k|k} \) will follow the same time update and measurement update equations \cite{41} and \cite{5}, respectively. However, the \( \gamma_k \) in \cite{41} for \( k \geq \Gamma \) will follow \cite{5}. Note that, after the attack start point, i.e., \( k \geq \Gamma \), the defender does not change the Kalman filter. However, since the attacker replaces \( y_k \) by the fake data \( z_k \) from \( k \geq \Gamma \), the innovation signal \( \gamma_k \) automatically takes the form given in \cite{5}.

As studied in several works of literature, the distribution of the attack start point can be modelled as exponential distribution for continuous-time systems \cite{41, 42}. In \cite{41}, the authors collected empirical data from intrusion experiments and found that the attack start points are approximately exponentially
distributed. On the other hand, the authors formalized the semantics of attack trees and used them for the probabilistic timed evaluation of attack scenarios in [42]. Additionally, the authors studied various practical systems, including the famous Stuxnet attack [4], and derived the distribution of the attack start time to be exponential. Since the exponential and geometric distributions play analogous roles in the continuous and discrete time domains, respectively [43], we have modelled the attack start point \( \Gamma \) to be an RV that follows a geometric distribution with parameter \( \rho \), where \( 0 < \rho < 1 \). Here \( \rho \) is a design parameter reflecting the defender’s belief of how often attacks occur. For the proposed method, \( \rho \) is a parameter that needs to be set by the defender. A vulnerability analysis of the system can decide the value of \( \rho \), see [41, 42].

From the derived approximate analytical expressions of ADD [60], FAR [61] and \( \Delta LQG \) [72], we can say that ADD and \( \Delta LQG \) will not be affected much by the difference in the chosen \( \rho \) and the attacker’s true \( \rho \), as long as both \( \rho \ll 1 \), which is a realistic assumption since attacks are rare events. On the other hand, FAR will increase if the chosen \( \rho \) is higher than the attacker’s true \( \rho \) and vice versa. The defender can thus choose a suitable \( \rho \), depending on the specification on the maximum false alarm rate. Finally, we can write the prior probability \( \Pi_k \triangleq \mathbb{P} \{ \Gamma = k \} \) in the following form [34],

\[
\Pi_k = \Pi_0 I_{\{k=0\}} + (1 - \Pi_0) \rho (1 - \rho)^{k-1} I_{\{k\geq 1\}}. \tag{13}
\]

Here, \( \Pi_0 \triangleq \mathbb{P} \{ \Gamma \leq 0 \} \), i.e., \( \Pi_0 \) is the prior probability of the attack happening before the start of the observation. \( I_{\{\text{condition}\}} \) is the indicator function, which takes the value 1 if the condition is true, or 0 otherwise. In general, \( 0 \leq \Pi_0 < 1 \). However, for our problem formulation, we have taken \( \Pi_0 = 0 \). We assume that the defender does not know the exact value of \( \Gamma \), but knows the prior distribution of \( \Gamma \).

III. PROPOSED DETECTION STRATEGY

We perform the following hypothesis test to detect the presence of an attack.

\( H_0 \): No attack present

\( H_1 \): Attack present in the system

We parsimoniously add an iid watermarking signal, given as

\[
e_k \sim N(0, \Sigma_e), \tag{14}
\]

to the optimal LQG control input, \( u_k^* \), to improve the detectability of the attack, see [17]. The decision of adding or not adding the watermarking and the selection of hypothesis for the \( k \)-th time instant is controlled by the optimal policy \( u_k^* \). Here, the subscript \( d \) of \( u_k^* \) indicates that the optimal policy is derived using dynamic programming. The policy \( u_k \) selects the values of the following two control variables \( s_k \) and \( d_k \) at the \( k \)-th time instant,

\[
s_k = \begin{cases} 
0, & \text{No watermarking for (k+1)-th time instant,} \\
1, & \text{Watermarking added for (k+1)-th time instant.} 
\end{cases} \tag{15}
\]

\[
d_k = \begin{cases} 
0, & \text{Hypothesis } H_0 \text{ selected, process continues.} \\
1, & \text{Hypothesis } H_1 \text{ selected, process terminated.} 
\end{cases} \tag{16}
\]

\[
u_k = u_k^* + s_{k-1} e_k. \tag{17}
\]

Figure 3 illustrates the proposed watermarking and attack detection scheme with a schematic diagram of the system. The components enclosed inside the blue dotted rectangle are assumed to be located at a secure location.

Figure 3: Schematic diagram of the system with the proposed watermarking scheme.

A. Problem formulation

Our objective is to find the optimal policy \( u_k^* \) that minimizes the ADD for fixed upper bounds on FAR and ANW. First, we introduce the formal definitions of FAR, ADD and ANW as follows. The definitions of FAR and ADD are similar to [31].

**False alarm rate (FAR):** FAR is defined as

\[
FAR \triangleq \mathbb{P}^\Pi \{ \tau < \Gamma \} . \tag{18}
\]

Here, \( \mathbb{P}^\Pi \) indicates the average probability measure. \( \mathbb{P}^\Pi \{ \Omega \} = \sum_{k=1}^{\infty} \Pi_k P_k \{ \Omega \} \), where \( \Omega \) is any event and \( P_k \) is the probability measure when the change point \( \Gamma = k \). \( \tau \) is the time instant when the hypothesis \( H_1 \) is selected.

**Average detection delay (ADD):** ADD is defined as

\[
ADD \triangleq \mathbb{E}^\Pi [ \tau - \Gamma | \tau \geq \Gamma ] . \tag{19}
\]

Here, \( \mathbb{E}^\Pi \) denotes the expectation with respect to the probability measure \( \mathbb{P}^\Pi \).

**Average number of watermarking events (ANW) before attack:** ANW is defined as

\[
ANW \triangleq \mathbb{E}^\Pi \left[ \sum_{i=1}^{\min(\tau, \Gamma - 1)} s_i \right] . \tag{20}
\]

Here, \( s_i \) is the same variable as given in (15).

Now, we formulate the following optimization problem,

\[
\min_{u_k} ADD, \tag{21}
\]

s.t. \( FAR \leq FAR_{th} \), \( ANW \leq ANW_{th} \),

where \( FAR_{th} \) and \( ANW_{th} \) are the user-selected thresholds. Then, the constrained optimization problem of (21) is converted into an unconstrained Lagrangian form as follows. The unconstrained Lagrangian form adopted in this paper is similar to [34], [44], except the term \( \lambda_e ANW \), and it reads as

\[
J^* = \min_{u_k} ADD + \lambda_f FAR + \lambda_e ANW, \tag{22}
\]

where \( \lambda_f \geq 0 \) and \( \lambda_e \geq 0 \) are the Lagrangian multipliers. A new state variable \( \theta_k \) is defined as

\[
\theta_k = \begin{cases} 
0, & \text{No attack,} \\
1, & \text{System under attack,} \\
T_c, & \text{Attack detected by hypothesis testing} \\
& \text{and process terminated.} 
\end{cases} \tag{23}
\]
Similar to [34], ADD, FAR and ANW can also be expressed in terms of the control variables, $s_k$ and $d_k$, and the state variable $\theta_k$ as follows,

$$ ADD = E \left[ \sum_{k=1}^{T} \mathbb{1}_{\{ \theta_k=1 \}} \mathbb{1}_{\{ d_k=0 \}} \right], \quad (24) $$

$$ FAR = E \left[ \sum_{k=1}^{T} \mathbb{1}_{\{ \theta_k=0 \}} \mathbb{1}_{\{ d_k=1 \}} \right], \quad (25) $$

$$ ANW = E \left[ \sum_{k=1}^{T} \mathbb{1}_{\{ \theta_k=0 \}} \mathbb{1}_{\{ s_k=1 \}} \mathbb{1}_{\{ d_k=0 \}} \right]. \quad (26) $$

Using (24)-(26), the cost function of (22) can be expressed as

$$ J^* = \min_{u_d} E \left[ \sum_{k=1}^{T} g_k(\theta_k, s_k, d_k) \right], \quad (27) $$

where $g_k(\cdot)$ is the per stage cost, expressed as

$$ g_k(\theta_k, s_k, d_k) = \mathbb{1}_{\{ \theta_k \neq T_k \}} \left[ \mathbb{1}_{\{ \theta_k=1 \}} \mathbb{1}_{\{ d_k=0 \}} + \lambda_f \mathbb{1}_{\{ \theta_k=0 \}} \mathbb{1}_{\{ d_k=1 \}} + \lambda_c \mathbb{1}_{\{ \theta_k=0 \}} \mathbb{1}_{\{ s_k=1 \}} \mathbb{1}_{\{ d_k=0 \}} \right]. \quad (28) $$

Here, the first, second and third terms of (28) come from ADD, FAR and ANW, respectively. For the stochastic optimal control problem defined in (27), the state $\theta_k$ is not observable to the defender. Therefore, we replace the state $\theta_k$ by its sufficient statistics $p_k$. The sufficient statistics $p_k$, i.e., the posterior probability of attack at $k$-th time instant is defined as $p_k \triangleq P \{ \Gamma \leq k | \Psi_k \} = E \left[ \mathbb{1}_{\{ \theta_k=1 \}} | \Psi_k \right]$, where $\Psi_k \sim \mathcal{N}$. In our study, the innovation signal $\gamma_k$ is iid before and non-iid after the attack. From (26), the Shiryaev statistics $SR_k$ at the $k$-th instant in time for our problem formulation can be written as

$$ SR_k = \sum_{i=1}^{k} \prod_{j=i}^{k} \frac{L_j}{1 - \rho}, \quad (31) $$

where $i$ is the candidate change or attack start point, and $L_j$ is the likelihood ratio. The expression for $L_j$ is given in Lemma 1. After the change point, $SR_k$ increases on average. In the original Shiryaev procedure, a change is detected once $SR_k$ crosses a predefined threshold for the first time.

**Lemma 1.** The likelihood ratio $L_j$ used to derive the Shiryaev statistics [31] considering the joint distribution of the innovation signals (29) and (30) and the watermarking signal (14), takes the following form,

$$ L_j = \begin{cases} L_{a,j}, & j > i, \\ L_{b,j}, & j = i \end{cases}, \quad (32) $$

$$ L_{a,j} = \frac{f_{j,i} \left( \gamma_j | \{ \gamma_{j-1}, s_{j-1} \} \right)}{f_0(\gamma_j)}, \quad (33) $$

$$ L_{b,j} = \frac{f_{j,j} \left( \gamma_j | \{ \gamma_{j-1}, s_{j-1} \} \right)}{f_0(\gamma_j)}. \quad (34) $$

Here, $f_{i,j} (\cdot)$, $f_{j,j} (\cdot)$ and $f_0 (\cdot)$ denote the distributions for $j > i$, $j = i$ and $j < i$, respectively, and $e_{s,j} = e_j - e_{j-1} e_j$. $f_{i,j} (\cdot)$ and $f_0 (\cdot)$ take the following forms,

$$ f_{i,j} (\cdot) = \mathcal{N} \left( \mu_{i,j}, \Sigma_{1,j} | z_{j-1}, x_{j-1,j-1}, e_{s,j-1} \right), \quad (35) $$

$$ f_{j,j} (\cdot) = \mathcal{N} \left( \mu_{j,j}, \Sigma_{2,j} | x_{j-1,j-1}, e_{s,j-1} \right), \quad (36) $$

$$ f_0 (\cdot) = \mathcal{N} \left( 0, \Sigma_0 \right). \quad (37) $$

Here

$$ \mu_{1,j} = A_0 z_{j-1} - C (A + BL) x_{j-1,j-1} - CB s_{j-1} e_{k-1}, \quad (38) $$

$$ \mu_{2,j} = -C (A + BL) z_{j-1,j-1} - CB e_{s,j-1}, \quad (39) $$

$$ \Sigma_{1,j} = Q_n \quad (40) $$

$$ \Sigma_{2,j} = Q_z \quad (41) $$

$$ \Sigma_0 = CPC^T + R \quad (42) $$
Proof. The proof of Lemma 1 is provided in Appendix A. □

Equation (31) is same as the original Shiryaev statistics, where \( \Pi_0 \) is assumed to be 0, see (6.9) from [26]. However, the term \( \mathcal{L}_j \) in (31) is derived exclusively for the problem under study, where the test data is iid before the change point and non-iid with stationary distributions after the change point. Furthermore, Lemma 1 shows that the dependency of the test data \( \gamma_j \) on the previous values of \( \gamma \) and \( e_s \) from the time index 1 to \( j - 1 \) can be approximated as given in (33)-(39), where \( \gamma_j \) is only dependent on the immediate past values, i.e., at time index \( j - 1, \) \( j, \) and \( e_s \).

Remark 2. The likelihood ratios using the distributions of the innovation signal alone, say, \( \mathcal{L}_j \) and \( \mathcal{L}_{d,j} \) for \( j < i \) and \( j = i \), respectively, can be evaluated from Lemma 1 by using \( e_{s,j-1} = 0 \) in (35)-(36). Therefore, we can write \( \mathcal{L}_{c,j} = \mathcal{L}_{a,j} \) \( |e_{s,j-1} = 0 \) and \( \mathcal{L}_{d,j} = \mathcal{L}_{b,j} \) \( |e_{s,j-1} = 0 \).

We have applied the value iteration from [45] using sufficient statistics \( p_{k} \) to solve (27), which is an infinite horizon dynamic programming problem with a termination state. The relationship between the Shiryaev statistics \( SR_k \) and the posterior probability of attack \( p_k \) is given by, see (6.10) from [26],

\[
p_k = \frac{SR_k}{SR_k + 1/\rho}.
\]  

Lemma 2 provides the recursion formula of \( p_k \), which is used for the value iteration.

Lemma 2. The posterior probability of attack at the \( k \)-th time instant, \( p_k \), for the test data \( \gamma_k \) (iid [29] and non-iid [30]) and \( e_s \) [27], can be updated using the following recursion formula, when the attack start point is geometrically distributed with parameter \( \rho \),

\[
p_k = \frac{p_{k-1}\mathcal{L}_{c,k} + (1 - p_{k-1}) \rho \mathcal{L}_{d,k}}{(1 - \rho) (1 - p_{k-1}) + p_{k-1} \mathcal{L}_{c,k} + (1 - p_{k-1}) \rho \mathcal{L}_{d,k}},
\]

when \( s_{k-2} = 0 \), and

\[
p_k = \frac{p_{k-1}\mathcal{L}_{a,k} + (1 - p_{k-1}) \rho \mathcal{L}_{b,k}}{(1 - \rho) (1 - p_{k-1}) + p_{k-1} \mathcal{L}_{a,k} + (1 - p_{k-1}) \rho \mathcal{L}_{b,k}},
\]

otherwise. \hspace{1cm} (44)

Proof. Using (31), the recursion formula of the Shiryaev statistics \( SR_k \) is derived first, see (45). Then using (43) and (45), the recursion equations of (44) are derived.

\[
SR_k = \begin{cases} 
\frac{\mathcal{L}_{c,k}}{1 - \rho} SR_{k-1} + \frac{\mathcal{L}_{d,k}}{1 - \rho} & s_{k-2} = 0, \\
\frac{\mathcal{L}_{a,k}}{1 - \rho} SR_{k-1} + \frac{\mathcal{L}_{b,k}}{1 - \rho} & s_{k-2} = 1.
\end{cases}
\] \hspace{1cm} (45)

We use the following simplified notations to represent the recursion formula in (44), \( p_k = \phi_0 (p_{k-1}) \), if \( s_{k-2} = 0 \), and \( p_k = \phi_1 (p_{k-1}) \), otherwise. Also, the initial value of \( p_k \) is taken to be 0.

3) Solution of optimization problem (27)

The expected value of the per stage cost function \( g_k(\cdot) \) in (27) is derived by taking expectations on both sides of (28), and using \( p_k = E \left[ \mathbb{1}_{\{d_k=0\}} | \Psi_k \right] \), see (46).

\[
E[g_k (\theta_k, s_k, d_k) | \Psi_k] = p_k \mathbb{1}_{\{d_k=0\}} + \\
\lambda_f (1 - p_k) \mathbb{1}_{\{d_k=1\}} + \lambda_e (1 - p_k) \mathbb{1}_{\{s_k=1\}} \mathbb{1}_{\{d_k=0\}}.
\] \hspace{1cm} (46)

The Bellman equation for the infinite horizon cost function (27) with the termination state \( T_c \) can be formulated using the sufficient statistics \( p_k \) as follows,

\[
J (p_k) = \min_{u_k \in \mathcal{U}} \left[ p_k \mathbb{1}_{\{d_k=0\}} + \lambda_f (1 - p_k) \mathbb{1}_{\{d_k=1\}} + \\
+ \lambda_e (1 - p_k) \mathbb{1}_{\{s_k=1\}} \mathbb{1}_{\{d_k=0\}} + B_0 (p_k) \mathbb{1}_{\{s_k=0\}} \mathbb{1}_{\{d_k=0\}} + \\
+ B_1 (p_k) \mathbb{1}_{\{s_k=1\}} \mathbb{1}_{\{d_k=0\}} \right],
\] \hspace{1cm} (47)

where \( \mathcal{U} = \{1, 2, 3\} \) is the set of all stationary deterministic permissible policies, see Table II: \( B_0 (p_k) = E \left[ J (\phi_0 (p_k)) \right] \), and \( B_1 (p_k) = E \left[ J (\phi_1 (p_k)) \right] \). Therefore, \( B_0 (p_k) \) and \( B_1 (p_k) \) denote the expected total costs from \((k + 1)\)-th time instant till the termination of the process when \( s_k = 0 \) and \( s_k = 1 \), respectively, and \( d_k = 0 \). Note that, when \( d_k = 1 \), the process terminates immediately, so there will be no additional cost after the \( k \)-th time instant. In addition to that, if \( d_k = 1 \) then the process will immediately terminate, and there will be no use of adding watermarking at the \((k + 1)\)-th time instant, therefore, the combination \((s_k = 1, d_k = 1)\) has been ignored.

| Table II: \( \mathcal{U} \) |
|------------------|
| \( u_{d,k} \) | \( s_k \) | \( d_k \) |
| 1 | 0 | 0 |
| 2 | 1 | 0 |
| 3 | 0 | 1 |

To satisfy the accessibility hypothesis, we have discretized the space of the sufficient statistics \( p_k \) into a finite set during the numerical simulations. On the other hand, the control space of the stochastic optimization problem under study is inherently discrete and finite. Finally, the value iteration is used to solve the Bellman equation (47) and to find the optimal policy \( u^*_d \).

C. Structural properties of the optimal policy

In this subsection, we will study the structure of the optimal solution found by solving the Bellman equation (47).

Assumption A.1: There exist at least one stationary deterministic policy \( u_d \), for which both the constraints as given in (21) will be satisfied.

Assumption A.1 is about the feasibility of the existence of a stationary deterministic policy for the optimization problem in (21). Now the value iteration reveals the following optimal policy for selecting the control variables \( s_k \) and \( d_k \) values.

\[
s_k = \begin{cases} 
0, & B_0 (p_k) - B_1 (p_k) < \lambda_e (1 - p_k), \\
1, & \text{otherwise}.
\end{cases}
\] \hspace{1cm} (48)

\[
d_k = \begin{cases} 
0, & p_k + \lambda_e (1 - p_k) \mathbb{1}_{\{s_k=1\}} + B_0 (p_k) \mathbb{1}_{\{s_k=0\}} \\
+ B_1 (p_k) \mathbb{1}_{\{s_k=1\}} < \lambda_f (1 - p_k), \\
1, & \text{otherwise}.
\end{cases}
\] \hspace{1cm} (49)

First, we will prove that the optimal policy is going to be a stationary deterministic policy. As stated in Lemma 3.1 from [44], the costs FAR and ANW will be monotone and non-increasing in \( \lambda_f \) and \( \lambda_e \), respectively. We can prove that by following the similar steps used to prove Lemma 3.1 in [44]. From the monotone and non-increasing properties of FAR and ANW, it can be proved that the inequality conditions in the original constrained optimization problem (21) will be satisfied.
for finite values of $\lambda_f \geq 0$ and $\lambda_e \geq 0$ for some deterministic policy as stated in Lemma 3.3 from [44].

Finally, as discussed in [44], under assumption A.1 or the weaker condition of Lemma 3.1, the stationary deterministic optimal policy found by solving the Bellman equation (47) from the unconstrained optimization problem with the Lagrangian multipliers, $\lambda_e$ and $\lambda_f$, will be the solution of the original constrained problem as given in (21).

Even though a formal proof is unavailable at this point, we have performed extensive numerical simulations and found that for the following properties of the optimal policy: The optimal policy is a two threshold policy, $Th^*$ and $Th^d$, $Th^d \geq Th^*$, Figures 4 and 5 provide the insights with thresholds of a two-threshold policy as stated in Lemma 3.3 from [44]. The following steps can be followed offline to find the optimal policy found by solving the Bellman equation (47) from [34].

Figures 4 and 5 provide the insights with thresholds of a two-threshold policy by plotting the left-hand sides (LHS) and right-hand sides (RHS) of (48) and (49), respectively, for a relatively small $\lambda_e$ and large $\lambda_f$. We observe that for (48), the LHS crosses the RHS at two points, but the second crossing happens after $d_k = 1$, i.e., the termination of the process, which results in a one-two-threshold policy. We have found that for a relatively large $\lambda_e$ or $\rho$ near to unity, the optimal policy may even become a one or three-threshold policy, which is similar to the findings of [34]. The following steps can be followed offline to find the two thresholds.

Step 1: The search space of $\lambda_e$ and $\lambda_f$ is divided into $N$ equally spaced grid points.

Step 2: For each grid point, we perform the value iterations using (47), and store $J^*(p_k)$ where $p_k$ is also discretized in $[0, 1]$.

Step 3: For each grid point, ADD, FAR and ANW are evaluated from Monte-Carlo simulations by deriving the decision variables $s_k$ and $d_k$ from (48)-(49) using $J^*(p_k)$ from Step 2.

Step 4: Select the best $\lambda_e^*$ and $\lambda_f^*$ combination, which gives minimum ADD and satisfies the constraints on FAR and ANW, see (21).

Step 5: Apply numerical solvers such as the Trust-Region algorithm, the bisection method, etc., to solve the following two equations for $\bar{p}$, see (50) and (51). The solutions of (50) are $\bar{p} = Th^*$ and $\bar{p} = 1$. Equation (51) is derived from (49) using $s_k = 0$ and solved for $\bar{p} \in [Th^*, 1]$. The solutions of (51) is $\bar{p} = Th^d$ and $\bar{p} = 1$.

\begin{align}
B_0(\bar{p}) - B_1(\bar{p}) &= \lambda_e (1 - \bar{p}) \quad (50) \\
\bar{p} + \lambda_e (1 - \bar{p}) + B_1(\bar{p}) &= \lambda_f (1 - \bar{p}) \quad (51)
\end{align}

Finally, the optimal policy $u^*_d(k)$ is given as

\begin{align*}
u^*_d(k) &= \begin{cases} 1, & \text{i.e., } (s_k = 0, d_k = 0) \quad p_k < Th^*, \\
2, & \text{i.e., } (s_k = 1, d_k = 0) \quad p_k \geq Th^*, \\
3, & \text{i.e., } (s_k = 0, d_k = 1) \quad p_k \geq Th^d.
\end{cases}
\end{align*}

Next, we briefly discuss the computational runtime complexity of the proposed policy.

D. Computational complexity

The proposed technique is an online method. At run time, we only need to evaluate $p_k$ [44] and compare it with two thresholds 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 (32) can be derived offline since the variances are fixed, see (40)-(42). The most expensive operations at runtime are a few matrix-vector multiplications with the highest computational complexity of $O(np)$, see (38) and (39).

IV. DERIVATIONS OF ADD, FAR AND $\Delta LQG$

This section derives the asymptotically approximate analytical expressions of ADD, FAR and $\Delta LQG$ for the given thresholds $Th^*$ and $Th^d$, and a few other parameters to be defined later.

A. Approximate Expressions of ADD and FAR

Here we derive the approximate expressions of ADD and FAR, as $Th^d \to \infty$, applying non-linear renewal theory [31], [46]. First, the Shiryaev statistics $SR_k$ is converted into $LSR_k = \log (SR_k)$ for the ease of asymptotic analysis. $LSR_k$ can be expressed as a summation of two variables, $S_k$ and $l_k$, as given in the following Lemma 3.

**Lemma 3.** The logarithm of the Shiryaev statistics, $LSR_k$, generated from the test data, i.e., the innovation signal $\gamma_k$ ([29] and [37]) and the watermarking signal $e_k$ ([44], under the two threshold policy $Th^*$ and $Th^d$, can be expressed as the summation of two variables $S_k$ and $l_k$, see (53). Here $S_k$ ([54] is a ladder variable, and $l_k$ ([55] is a slowly changing variable in the sense defined in [46].

\begin{align}
LSR_k &= S_k + l_k. \quad (53)
\end{align}

\begin{align}
S_k &= Z_k + k|\log(1 - \rho)|. \quad (54)
\end{align}
\[ l_k = \log \left( SR_0 + \sum_{i=1}^{k} (1 - \rho)^{i-1} L_{d,i} \exp(-\lambda_i) \mathbb{I}_{\{\text{LSR}_i < \text{Th}^s\}} \right) + \sum_{i=1}^{k} (1 - \rho)^{i-1} L_{b,i} \exp(-\lambda_i) \mathbb{I}_{\{\text{LSR}_i \geq \text{Th}^s\}} \right) + \sum_{j=1}^{k} \log (L_{c,j}) \mathbb{I}_{\{\text{LSR}_j < \text{Th}^s\}} - \sum_{j=1}^{k} \log (L_{a,j}) \mathbb{I}_{\{\text{LSR}_j < \text{Th}^s\}}, \]  

where

\[ Z_k = \sum_{i=1}^{k} \log (L_{a,i}), \]  

\[ \lambda_k = \sum_{i=1}^{k} \log (L_{a,i}) \mathbb{I}_{\{\text{LSR}_i \geq \text{Th}^s\}} + \sum_{i=1}^{k} \log (L_{c,i}) \mathbb{I}_{\{\text{LSR}_i < \text{Th}^s\}}, \]  

and

\[ \text{Th}^s = \log \frac{\text{Th}^d}{\rho (1 - \text{Th}^d)}. \]  

Proof. The proof of Lemma [3] is provided in Appendix E. \( \square \)

Remark 3. The threshold \( \text{Th}^s \) for \( p_k \) is equivalent to the threshold \( \text{Th}^d \) for LSRk. Similarly, we can define the threshold \( \text{Th}^D \) as

\[ \text{Th}^D \equiv \log \frac{\text{Th}^d}{\rho (1 - \text{Th}^d)} \]  

for LSRk, which is equivalent to the threshold \( \text{Th}^d \) for \( p_k \).

Therefore, Lemma [3] enables us to apply non-linear renewal theory to derive the approximate expressions of ADD and FAR by splitting the logarithm of the Shiryaev statistics, LSRk, into a ladder variable \( S_k \) and a slowly changing term \( l_k \). The definition of a slowly changing variable from [46] is also provided in Appendix E. We define the variable \( r \) to be the overshoot of the ladder variable \( S_{n_d} \) over a large threshold \( \text{Th}^D \) at \( k = n_d \). Therefore, \( r_{n_d} \equiv S_{n_d} - \text{Th}^D \) as \( \text{Th}^D \to \infty \), and \( n_d = \inf \{ k \geq 1 : S_k \geq \text{Th}^D \} \). According to the non-linear renewal theory [46], the overshoot statistics of LSRk crossing a large threshold \( \text{Th}^D \) can be approximated as the statistics of \( r_{n_d} \), provided \( l_k \) is slowly changing and \( \text{Th}^D \to \infty \). The approximate expressions of ADD and FAR derived in this paper are stated in Theorem 1.

Theorem 1. For the Shiryaev statistics given in Lemma [7] and the geometric prior distribution of the change point \( \Gamma \) [13], under the two threshold policy \( \text{Th}^s \) and \( \text{Th}^d \), the asymptotic approximate expressions of ADD and FAR as \( \text{Th}^D \to \infty \) will take the following forms, provided the conditions C1-C4 are satisfied.

Conditions:

C1: \( \{ Z_k : k \geq 1 \} \) is non-arithmetic with respect to \( P_0 \) and \( P_1 \).

C2: \( E_1 [ Z_1 ] \) is finite.

C3: \( l_k \) is a slowly changing variable in the sense defined in [46].

C4: \( 0 < E_1 [ D (f_1^*, f_0)] \) < \( \infty \), and \( 0 < E_0 [ D (f_0, f_1^*)] \) < \( \infty \).

Then,

\[ \text{ADD} = \frac{\text{Th}^D + \bar{v} - \bar{i}}{E_1 [ D (f_1^*, f_0)] + \| \log(1 - \rho) \| + o(1), \]  

and \( \text{FAR} \approx \frac{\xi}{\rho \exp (\text{Th}^D)} (1 + o(1)), \) as \( \text{Th}^D \to \infty \),

where

\[ \bar{v} = \lim \frac{E_1 [ r_{n_d}]}{n_d \to \infty}, \]  

\[ \bar{i} = \lim \frac{E_1 [ l_k]}{k \to \infty}, \]  

\[ \xi = \lim \frac{E_1 [ \exp (-r_{n_d})]}{n_d \to \infty}. \]  

\( P_0 \) and \( P_1 \) denote the probability measures before and after the attack, respectively. \( E_0 \) and \( E_1 \) denote the expectations with respect to the probability measures \( P_0 \) and \( P_1 \). \( E_1 [ D (f_1^*, f_0)] \) is the expected KLD between the distributions \( f_{1,j}^* (\cdot) \) and \( f_{2,j} (\cdot) \), and \( f_{1,j} (\cdot) = f_{1,j} (\cdot) \) when \( s_j = 1 \) for all \( j \). Here, the expectation is taken over the joint distribution of the innovation signal and the watermarking signal after the attack start point. Similarly, \( E_0 [ D (f_0, f_1^*)] \) is the expected KLD between \( f_0 \) and \( f_1^* \), and the expectation is taken over the joint distribution of the innovation signal and the watermarking signal before the attack start point.

Proof. The proof of Theorem 1 is provided in Appendix C. \( \square \)

Remark 4. As given in Lemma 1 and Theorem 2 of [59], \( E_1 [ D (f_1^*, f_0)] \) will take the following form,

\[ E_1 [ D (f_1^*, f_0)] = \frac{1}{2} \left\{ \text{tr} (\Sigma_0^{-1} \Sigma_{\gamma}) - m - \log \frac{Q_{\gamma}}{\Sigma_0} \right\}, \]  

where the covariance matrix \( \Sigma_{\gamma} \) of the innovation signal after the attack start point is given as

\[ \Sigma_{\gamma} = E_{zz} (0) - C (A + BL) E_{zz} (1) - [C (A + BL) E_{zz} (1)]^T + CB \Sigma_{zz} B^T C^T + C (A + BL) \Sigma_{s \times e} (A + BL)^T C^T + C (A + BL) \Sigma_{s \times e} (A + BL)^T C^T, \]  

where \( E_{zz} (1) = \sum_{i=0}^{\infty} A^i K A^{i+1} E_{zz} (0) \) and \( E_{zz} (0) = E [ z_k z_k^T ] \). \( \Sigma_{s \times e} \) and \( \Sigma_{s \times e} \) are the solutions to the following Lyapunov equations,

\[ A \Sigma_{s \times e} A^T - \Sigma_{s \times e} + KE_{zz} (0) \right) K^T + AE_{zz} (1) K^T \]  

= \( 0 \), and

\[ A \Sigma_{s \times e} A^T - \Sigma_{s \times e} + (I_n - KC) B \Sigma_{zz} B^T (I_n - KC)^T = 0. \]  

Here \( A = (I_n - KC) (A + BL) \), which is assumed to be strictly stable. \( I_n \) is an identity matrix of size \( n \times n \).

Therefore, we can derive approximate values of ADD and FAR using Theorem 1 for the given thresholds, \( Th^D \) and \( Th^s \), and the system and noise parameters. The denominator of [60] does not depend on the thresholds. Also, according to the renewal theory, the statistics obtained from the overshoot \( r_{n_d} \), i.e., \( \bar{v} \) and \( \xi \), are not dependent on the exact values of the thresholds as long as \( \text{Th}^d \) is large enough. However, from [55],
we can say that $\bar{l}$ is dependent on the threshold $Th^s$. Further approximation of the expression of ADD (60) can be directly obtained from Theorem 1 as stated in Corollary 1.1.

**Corollary 1.1.** The approximate expression of ADD as provided in Theorem 1 can be further simplified as follows,

$$ADD \approx \frac{Th^D}{E_1[D(f^c_s,f_0)] + \log(1-\rho)}$$

as $Th^D \to \infty$ (70)

**Proof.** $\bar{r} \ll Th^D$ and $\bar{l} \ll Th^D$, since $Th^D \to \infty$. Therefore, by ignoring $\bar{r}$ and $\bar{l}$ from (60), we get (70).

The approximate expression of ADD as provided in Corollary 1.1 does not depend on the threshold $Th^s$. Therefore, Corollary 1.1 can also be used to find a suitable value of the threshold $Th^D$ for a given ADD.

**Remark 5.** Finding analytical expressions for $\bar{r}$, $\bar{l}$, and $\xi$ is difficult for the system under consideration. Therefore, we estimate their values by Monte-Carlo (MC) simulation. The values of $\bar{r}$, $\bar{l}$, and $\xi$ are not directly dependent on $Th^D$ as long as $Th^D$ is very large, but they depend on $Th^S$. However, to derive the ADD using the second approximate expression as given in Corollary 1.1, we do not need the values of $\bar{r}$, $\bar{l}$, and $\xi$, but it is less accurate compared to Theorem 1.

**Remark 6.** In order to use the quickest detection scheme, the pre and post-change pdfs must be known. To achieve that, we need to know $A_a$ and $Q_a$. In practice, it is highly likely that the attacker’s system parameters $A_a$ and $Q_a$ may not be known a priori. In such a case, the attacker’s system parameters $A_a$ and $Q_a$ can be estimated online from the received observations (true or fake) by fitting a vector autoregressive model to the observations (47). This estimator will operate in parallel with the attack detection algorithm. Such a parameter estimation scheme will operate before and after the attack. However, before the attack, the estimates of $A_a$ and $Q_a$ will represent the healthy plant model. We have conducted some preliminary studies using a MISO system where the attack detection algorithm can perform with estimated parameters, albeit with additional watermarking compared to the known parameter case. A detailed analysis of such a joint estimation and detection scheme is however, beyond the scope of the current manuscript, and interested readers are referred to [48]. Additionally, we also comment that under a replay attack, $A_a$ and $Q_a$ can be derived from the normal system model as discussed in (27).

**B. Approximate Expression of ANW and $\Delta LQG$**

Following similar steps as in (34), the ANW can be approximated as follows,

$$ANW \approx \frac{E_0 \left[ t_1 \left( Th^S \right) \right]}{E_0 \left[ t_1 \left( Th^S \right) \right] + E_0 \left[ t_2 \left( LSR_{T\bar{r}h^S}, Th^S \right) \right]} \times P \left\{ t \left( Th^S \right) < \Gamma \right\}$$

(71)

Here, $t_1 \left( Th^S \right)$ denotes the time interval between the time instances when $LSR_k$ starts from $Th^S$, and then crosses the threshold $Th^S$ from above. $t_2 \left( LSR_{T\bar{r}h^S}, Th^S \right)$ denotes the time interval between the time instances when $LSR_k$ starts from $LSR_{T\bar{r}h^S}$ and crosses the threshold $Th^S$ from below. $t \left( Th^S \right)$ is the first time $LSR_k$ crosses the threshold $Th^S$ from below. An example plot of $LSR_k$ is shown in Fig. 6 to illustrate the variables $t_1(\cdot)$, $t_2(\cdot)$, and $t(\cdot)$. $E_0[\cdot]$ denotes the expectation with respect to the probability measure before the attack. Deriving analytical expressions for the expectations and the probability values in (71) is difficult. Therefore, we perform MC simulation to estimate the ANW for the given thresholds $Th^S$ and $Th^D$. The relationship between the ANW and the increase in the control cost is given in the following theorem.

**Theorem 2.** For the parsimonious watermarking scheme adopted in this paper, the increase in the LQG control cost, $\Delta LQG$, is related to ANW as

$$\Delta LQG = \rho ANW tr \left( H \Sigma_e \right)$$

where

$$H = B^T \Sigma_L B + U$$

and $\Sigma_L$ is the solution to the following Lyapunov equation.

$$(A + BL)^T \Sigma_L (A + BL) - \Sigma_L + L^T UL + W = 0$$

(74)

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

Theorem 2 shows that $\Delta LQG$ is proportional to ANW and a linear function of the watermarking signal variance $\Sigma_e$. If watermarking is added at all the time instants, $\rho ANW$ will become unity, and Theorem 2 will coincide with the special case of always present watermarking as stated in Theorem 3 in [39].

**C. Comparative Analysis**

The proposed method is compared with the following two methods, PW-$\Sigma_e$: persistent watermarking with fixed watermarking power and PW-$\Delta LQG$: persistent watermarking with fixed $\Delta LQG$. The only difference between the proposed method and PW-$\Sigma_e$ is that the watermarking is always present for the latter, and the watermarking power for both the methods is $\Sigma_e$. On the other hand, the only difference between the proposed method and PW-$\Delta LQG$ is that the watermarking is always present for the latter, and the $\Delta LQG$ value is the same for both. The subscripts $P$, $A$ and $B$ denote the proposed method, PW-$\Sigma_e$, and PW-$\Delta LQG$, respectively.

1) **Comparison with PW-$\Sigma_e$**

**Claim 1.** The proposed optimal watermarking policy incurs a lesser increase in LQG cost compared to PW-$\Sigma_e$.

The increase in the LQG control cost for PW-$\Sigma_e$ is as follows, see Theorem 3 from [39].

$$\Delta LQG_A = tr \left( H \Sigma_e \right)$$

(75)
By comparing the increase in the LQG control cost between the two methods, we can write
\[
\Delta LQG = \Delta LQG_A - \Delta LQG_P = (1 - \rho ANW) \Delta LQG_A,
\]
where \(\Delta LQG_P\) denotes the increase in the LQG control cost for the proposed method. Since, \((1 - \rho ANW) < 1\), we can make Claim 2.

**Claim 2.** The increase in ADD for the proposed optimal policy with respect to PW-\(\Sigma_e\) will be small.

There will be an increase in the ADD for the proposed method. Other than the mean of the slowly changing term, \(\bar{I}\), in the ADD expression (60), the rest of the components will be the same for the proposed method and PW-\(\Sigma_e\). Therefore, the increase in the ADD for the proposed method will be as follows,
\[
\Delta ADD = ADD_P - ADD_A \approx \frac{\bar{I}_A - \bar{I}_P}{E_1[D(f_i^e, f_0)] + |\log(1 - \rho)|},
\]
where the subscripts \(A\) and \(P\) denote the PW-\(\Sigma_e\) and the proposed method, respectively. \(o(1)\) notation is dropped for simplicity. Here \(\bar{I}_P\) is the same as given by (55) and (63), and \(\bar{I}_A\) will take the following form
\[
\bar{I}_A = \lim_{k \to \infty} E_1[|A_{A,k}|], \quad \text{and} \quad l_{A,k} = \log \left( S R_0 + \sum_{i=1}^{k} (1 - \rho)^{-1} L_{b,i} \exp(-Z_i) \right).
\]
Since, \(\bar{I}_A\) and \(\bar{I}_P\) both are small quantities compared to \(Th^D\), which is assumed to be \(\to \infty\), we can make Claim 2.

**Claim 3.** The FAR for the proposed optimal policy and PW-\(\Sigma_e\) will almost be the same.

Since \(Th^D \geq Th^S\), the watermarking will be present for both cases when \(Z_n\) crosses the threshold \(Th^D\). In other words, the statistics of the overshoot \(r_{n,q}\) will be the same for both methods. Therefore, we can make Claim 3.

2) Comparison with PW-\(\Delta LQG\)

**Claim 4.** The watermarking signal power for the proposed optimal policy, \(\Sigma_{e,P}\), will be greater than or equal to the watermarking signal power of PW-\(\Delta LQG\), \(\Sigma_{e,B}\).

Since the increase in the LQG control cost is taken to be the same for both the methods, the watermarking signal powers \(\Sigma_{e,B}\) and \(\Sigma_{e,P}\) for the method PW-\(\Delta LQG\) and the proposed method, respectively, will be different. The relationships between the watermarking signal power and \(\Delta LQG\) for both the methods are given as,
\[
\Delta LQG = tr(H\Sigma_{e,B}) = \rho ANW tr(H\Sigma_{e,P}).
\]
Since \(\rho ANW \leq 1\), from (79) we can make Claim 4.

**Claim 5.** The ADD for the proposed optimal policy will be less than or equal to the ADD for PW-\(\Delta LQG\).

We use the ADD expression from Corollary 1.1 to compare the two methods. The difference in the ADD will be due to the difference in \(E_1[D(f_i^e, f_0)]_B\) and \(E_1[D(f_i^e, f_0)]_P\), as follows.
\[
E_1[D(f_i^e, f_0)]_P - E_1[D(f_i^e, f_0)]_B \\
\approx \frac{1}{2} tr(\Sigma^{-1}_\gamma (\Sigma_{\gamma_P} - \Sigma_{\gamma_A})).\]

Here the subscripts \(B\) and \(P\) denote the method PW-\(\Delta LQG\) and the proposed method, respectively. By examining (66), we can say \(\Sigma_{\gamma_P} - \Sigma_{\gamma_A} \geq 0\). Therefore, from (80), we can write \(E_1[D(f_i^e, f_0)]_P - E_1[D(f_i^e, f_0)]_B \geq 0\), and we can further make Claim 5.

**D. Optimum \(\Sigma_e\)**

Theorem 1 and Corollary 1.1 imply that the increase in KLD will reduce ADD. Therefore, we derive the optimum \(\Sigma_e\) that will maximize KLD for a given fixed upper limit on \(\Delta LQG\), denoted as \(\Delta LQG_P\) for the proposed method. The optimization problem is defined as follows.
\[
\max_{\Sigma_e} E_1[D(f_i^e, f_0)], \quad \text{s.t.} \quad \Delta LQG_P \leq J, \quad \Sigma_e \geq 0,
\]
where \(J\) is a user-defined threshold. As given in Remark 4, the KLD expression for the proposed parsimonious watermarking policy is identical with the case where watermarking is always present, i.e., the method PW-\(\Sigma_e\). Moreover, \(\Delta LQG_P\) (72) is just a scaled version of \(\Delta LQG_A\) (75). Therefore, the condition \(\Delta LQG_P \leq J\) in (81) can be replaced by \(\Delta LQG_A \leq J_A\), where \(J_A = J/(\rho ANW)\), without any change in the optimum \(\Sigma_e\) value. Now, the optimization problem for the proposed method becomes identical to the optimization problem for the method PW-\(\Sigma_e\). According to Theorem 4 from [39], the optimum \(\Sigma_e\) for PW-\(\Sigma_e\) will be a rank one positive semi-definite matrix. Therefore, the optimization problem in (81) can be written as
\[
\max_{\nu_\lambda} E_1[D(f_i^e, f_0)], \quad \text{s.t.} \quad \Delta LQG_A \leq J_A,
\]
where \(\nu_\lambda = \sqrt{\sigma_e} v_\lambda\), \(\sigma_e\) is the non-zero eigenvalue of \(\Sigma_e\) and \(v_\lambda\) is the corresponding eigenvector. As discussed in [39], the maximization of \(E_1[D(f_i^e, f_0)]\) with respect to \(\nu_\lambda\) is same as maximizing the following function,
\[
\max_{\nu_\lambda} \nu_\lambda^T H_{KLD} \nu_\lambda, \quad \text{s.t.} \quad \Delta LQG_A \leq J_A.
\]
where
\[
H_{KLD} = B^T (I_n - KC)^T \kappa_e (I_n - KC) B + B^T C^T C B.
\]
Here, \(\kappa_e\) is the solution to the Lyapunov equation
\[
A^T \kappa_e A - \kappa_e + (A + BL)^T C^T C (A + BL) = 0.
\]
Since the matrix \(A\) is assumed to be strictly stable, the Lyapunov equation of (85) will have a unique solution. As discussed in [39], the optimization problem of (85) can be solved by various methods available in the literature, such as sequential quadratic programming (SQP) [49], interior point method [50], simple gradient-based method [39], etc. Interested readers are referred to [39] for a detailed analysis, but the same has been removed from the current paper due to space constraints.
V. NUMERICAL RESULTS

This section will illustrate and validate different aspects of the proposed methodology using, System-A: a second-order multi-input single-output (MISO) open-loop unstable system and System-B: a fourth-order MIMO open-loop stable system. Appendix F provides the required parameters for simulations associated with System A and B.

1) Optimal Policy

Figure 7 shows the optimal decision variable $u^*_d,k$ vs. $p_k$ plots for three different values of $\lambda_c$ and a fixed $\lambda_f$ for System-A. The watermarking signal variance is taken to be a diagonal matrix with equal signal power, $\sigma^2_e$. We observe that the optimal policy is a two threshold policy, which validates the theory presented in Sub-section III-C. A higher $\lambda_c$ means a stricter constraint on how much watermarking could be added, which gets reflected into higher $Th^e$. On the other hand, a higher $Th^e$ means watermarking will be added for fewer samples. As discussed in Sub-section IV-C1 since the added watermarking has little effect on the FAR, the change in $\lambda_c$ does not affect the threshold $Th^d$ much.

Figure 8: Optimal policy for different $\lambda_f$ and fixed $\lambda_c$ for System-A. $\sigma^2_e = 1.19$.

Figure 8 shows the optimal decision variable $u^*_d,k$ vs. $p_k$ plots for three different values of $\lambda_f$ and a fixed $\lambda_c$ for System-A. The watermarking signal variance is taken to be a diagonal matrix with equal signal power, $\sigma^2_e$. A higher $\lambda_f$ means a stricter constraint on how much FAR could be allowed. Therefore, the increase in $\lambda_f$ increases the threshold $Th^d$. However, since $Th^d \geq Th^e$, the change in $\lambda_f$ does not affect the threshold $Th^e$.

Figure 9: $s_k$, $d_k$ and $p_k$ vs. $k$ for a sample trial run for System-A. $\lambda_c = 0.2$, $\lambda_f = 100$ and $\sigma^2_e = 1.19$.

3) ADD and FAR vs. $\sigma^2_e$

Figure 10 shows the comparison between the plots of ADD and FAR vs. $\sigma^2_e$ for two different values of $\lambda_c$ for System-A. $\Sigma_e$ is taken to be a diagonal matrix with equal signal power, i.e., $\sigma^2_e$. For each $\sigma^2_e$ point, the thresholds $Th^e$ and $Th^d$ are derived using value iterations from dynamic programming. Then, the ADD and FAR are estimated by MC simulations using the derived thresholds. As discussed before, higher $\lambda_c$ reduces the usage of watermarking before the attack by increasing the threshold $Th^e$. The derived approximate expression of ADD (60) reveals that the ADD does not depend on $\lambda_c$ or $Th^e$ directly. However, from (55) and (63), we can say that $l$ reduces with the reduction in watermarking, which in turn increases ADD. Since $l$ is a small quantity compared to $Th^D$, the effect of the change of $l$ is small on ADD. To summarize, lower $\lambda_c$ results in slightly lower ADD. Similarly, the derived approximate expression of FAR (61) reveals that the FAR does not depend on $\lambda_c$ or $Th^e$ also. That is why we observe very similar FAR curves for two different values of $\lambda_c$ in Fig. 10.

Figure 10: ADD and FAR (%) vs. $\sigma^2_e$ plots for two different $\lambda_c$ and $\lambda_f = 100$, for System-A.

Figure 11 compares the same set of plots as in Figure 10 but for two different values of $\lambda_f$ and a fixed $\lambda_c$ for System-A. As discussed before, the increase in $\lambda_f$ increases $Th^d$. From (60) and (61), we know that ADD and FAR are mainly dependent on the value of $Th^D$. ADD increases with the increase in $Th^D$, power, $\sigma^2_e$. We have also indicated the attack start point as “change pt” in the plot. This figure provides relevant insights into how the proposed method works. We can observe that only for a very few time instances $p_k \geq Th^e$, and watermarking have been added before the attack. Such parsimonious use of watermarking reduces the control cost before the attack. On the other hand, $p_k$ increases gradually after the attack start point and eventually crosses the threshold $Th^d$. In other words, $p_k \geq Th^e$ and watermarking have been added almost all the time after the change point, resulting in faster detection.
whereas FAR reduces. To summarize, ADD increases and FAR decreases with $\lambda_f$.

In both the figures, Fig. 10 and Fig. 11 ADD reduces with the increase in the watermarking signal power, which is primarily the result of increased KLD (65). On the other hand, FAR (61) does not reduce much with the watermarking signal power since the correlation is weak. Higher watermarking signal power increases the overshoot $r_{nd}$ to some extent, which in turn reduces $\xi$ (64) slightly.

Figure 11: ADD and FAR (%) vs. $\sigma_e^2$ plots for two different $\lambda_f$ and $\lambda_e = 0.2$, for System-A.

4) ADD, FAR and $\Delta LQG$ theoretical values

Figure 12 shows the ADD and FAR vs. $\sigma_e^2$ plots for System-A, where ADD and FAR are estimated by MC simulations and also derived using Theorem 1 and Corollary 1.1. The watermarking signal variance is taken to be a diagonal matrix with equal signal power, i.e., $\sigma_e^2$. For each $\sigma_e^2$ point, the thresholds $Th^a$ and $Th^d$ are derived using dynamic programming value iterations. Figure 13 shows the same set of plots as in Fig. 12 for System-B. The ADD derived using MC simulations does not reduce at the same rate as that of the approximate theoretical ADD with the increase in $\sigma_e^2$. The reason is that the derived analytical expression of ADD is asymptotically approximate. On the other hand, we have selected the parameter values for the MC simulations so that ADD remains small for the ease of simulation studies. Within the small delay window after the change point for the MC simulations, the increase in $\sigma_e^2$ is not making much difference to the estimated ADD. On the other hand, the simulation study shows that $\xi$ (64) does not change much for a small increase in $\sigma_e^2$. Therefore, from (61), we can say FAR will only be affected to a small extent due to the increase in $\sigma_e^2$. Therefore, we observe that the simulated FAR and the theoretical FAR are in close agreement in Fig. 12. We also see that the derived ADD from Theorem 1 is a better match compared to the ADD derived from Corollary 1.1.

Figure 13: Comparison between the estimated values and theoretical values. ADD and FAR (%) vs. $\sigma_e^2$ plot for System-B. $\lambda_f = 100$ and $\lambda_e = 0.2$.

5) Comparison with PW-$\Sigma_e$

Figure 14 compares the $\Delta LQG$ vs. $\sigma_e^2$ plot, where $\Delta LQG$ is estimated by MC simulation and also derived using the theory presented in this paper for System-A using the same parameters as Fig. 12. The watermarking signal variance is taken to be a diagonal matrix with equal signal power, i.e., $\sigma_e^2$. From the derived expression of $\Delta LQG$ (72), it is evident that the control cost will increase with the increase in watermarking signal power.

Figure 14: Comparison between the estimated values and theoretical values. $\Delta LQG$ vs. $\sigma_e^2$ plot for System-A. $\lambda_f = 100$ and $\lambda_e = 0.2$.

We have shown ADD and FAR vs. $\sigma_e^2$ plots for the proposed method and PW-$\Sigma_e$ assuming a diagonal $\Sigma_e$ with equal power, $\sigma_e^2$, for System-A. For each $\sigma_e^2$ point, the thresholds $Th^a$ and $Th^d$ are derived using dynamic programming value iterations for the proposed method, and the same thresholds are used for PW-$\Sigma_e$ for a fair comparison. From the derived expression of $\Delta LQG$ (76), we predicted that we would get a large improvement in the control cost since $\rho$ and ANW both are small quantities. Also, the difference will increase with $\sigma_e^2$ as $\Delta LQG_A$ increases with $\Sigma_e$. As predicted from the theory discussed in Sub-section IV-C1, we observe a large improvement in the control cost (approx. 99% reduction in $\Delta LQG$) for the proposed method in Fig. 15 which validates our Claim 1.

Figure 15: Comparison between the estimated values and theoretical values. $\Delta LQG$ vs. $\sigma_e^2$ plot for System-A. $\lambda_f = 100$ and $\lambda_e = 0.2$.

From the derived expression of $\Delta ADD$ (77), we can comment that the proposed method will take a longer time on average to detect the attack compared to PW-$\Sigma_e$. The difference is due to the slowly changing terms, $\bar{l}_A$ and $\bar{l}_P$. Since the magnitude of the slowly changing term usually remains small, the increase in ADD for the proposed method is also small. In Fig. 16 an average increase of 35% (approx.) in ADD is observed at the same FAR for the proposed method. On the
To summarize, Fig. 16 supports our Claim 2 and Claim 3.

other hand, as discussed in Sub-section IV-C1, FAR will be the same for both the methods and the same is observed in Fig. 16. To summarize, Fig. 16 supports our Claim 2 and Claim 3.

Figure 16: Comparison between proposed method and PW-$\Sigma_e$. ADD and FAR vs. $\sigma_e^2$ plot for System-A. $\lambda_f = 100$ and $\lambda_e = 0.2$.

6) Comparison with PW-$\Delta LQG$

We have shown ADD and FAR vs. $\Delta LQG$ plots derived from MC simulations for the proposed method and PW-$\Delta LQG$ assuming a diagonal $\Sigma_e$ with equal power $\sigma_e^2$ in Fig. 17 for System-A. For each $\Delta LQG$ point, the thresholds $Th^s$ and $Th^d$ are derived using dynamic programming value iterations for the proposed method, and the same thresholds are used for PW-$\Delta LQG$ for a fair comparison. In general, for the proposed method, $Th^s$ decreases and $Th^d$ increases with the increase in $\Delta LQG$ or $\sigma_e^2$ for fixed $\lambda_e$ and $\lambda_f$. Since the same thresholds are used for PW-$\Delta LQG$, the ADD increases with $\Delta LQG$ in the plot. As discussed in Sub-section IV-C2 since the proposed method uses a higher watermarking signal variance at the same control cost, the KLD for the proposed method is higher compared to PW-$\Delta LQG$. Higher KLD for the proposed method results in lower ADD, and the same characteristic is observed in Fig. 17. Also, the usage of higher watermarking signal power increases the overshoot statistic to a small extent, resulting in a small decrease in FAR. To summarize, Fig. 16 supports our Claim 4 and Claim 5.

7) Optimum $\Sigma_e$

As discussed in Sub-section IV-D, the optimum $\Sigma_e^*$ reduces the KLD for a fixed upper bound on the $\Delta LQG$, which in turn reduces the ADD. We compare the ADD for the optimum $\Sigma_e^*$ and the diagonal $\Sigma_e$ in Fig. 18 for System-A. For each $\sigma_e^2$ point, the thresholds $Th^s$ and $Th^d$ are derived using dynamic programming value iterations for the diagonal $\Sigma_e$ case, and the same thresholds are used for the optimum $\Sigma_e^*$ case for a fair comparison. We observe an average increase of 14% (approx.) in the estimated ADD for the optimal $\Sigma_e^*$. For the optimal $\Sigma_e^*$, the watermarking signal power is mostly concentrated in one eigenvector direction, which results in higher overshoot and a lower FAR.

Figure 17: Comparison between proposed method and PW-$\Delta LQG$. ADD and FAR vs. $\Delta LQG$ plot for System-A. $\lambda_f = 100$ and $\lambda_e = 0.3$.

8) Comparison with a periodic watermarking scheme

We have compared the proposed evidence-based parsimonious watermarking scheme with a periodic watermarking scheme. The periodic watermarking scheme is adopted from [18] for our problem formulation. To fairly compare both methods, we have evaluated ADD and FAR by MC simulations for the same $\Delta LQG$ values. Under both schemes, the $p_k$ has been evaluated and compared with the same $Th^d$ value for attack detections. Note that $Th^d$ values are different for different $\Delta LQG$ values. However, watermarking has been added under the proposed scheme if $p_k \geq Th^s$. On the other hand, watermarking is added only once in a period for the other method, and the periods are determined separately for each $\Delta LQG$ value. Since the periodic watermarking scheme does use any existing evidence extracted from the set $\Psi_k$ of all available information up to the $k$-th time instant, the watermarking frequency remains the same before and after the attack. However, for the proposed scheme, the watermarking frequency increases significantly after the attack (approx. 50 times), which reduces ADD and FAR, see Fig. 19.

VI. CONCLUSION

In this paper, we have studied the quickest data deception attack detection problem with constraints on FAR and ANW. Such parsimonious use of watermarking helps to reduce the control cost during normal system operations and maintain a moderate detection performance. First, we have formulated the problem as a stochastic optimal control problem under a Bayesian framework. Then, we have applied dynamic programming to find the optimal policy. We have studied the optimal policy structure and found the optimal policy to be
a two threshold policy on the posterior probability of attack under a few practical assumptions. We have also derived the asymptotic approximate expressions of ADD and FAR applying non-linear renewal theory. The analytical expression of $\Delta LQG$ and its relationship with ANW is also derived. Theoretical and simulation studies reveal significant improvement in reducing $\Delta LQG$ with a relatively small increase in ADD compared to the method PW-$\Sigma_{eL}$, where watermarking is always present. The proposed method is also compared with PW-$\Delta LQG$, where both the methods have the same $\Delta LQG$ limit and found that the proposed method performs better in terms of ADD and FAR. Furthermore, we have described a technique to find the optimal watermarking signal power that maximises the KLD, which will improve the ADD.

**APPENDIX A**

**PROOF OF LEMMA 3**

The likelihood ratio, $L_{a,j}$, of the joint dependent distributions of the innovation signal and the watermarking signal, after and before the attack, takes the following form,

$$L_{a,j} = \frac{f_{a,j}(\gamma_j, e_{s,j-1}|\{\gamma\}_1^{j-1}, \{e_s\}_1^{j-2})}{f_{0,j}(\gamma_j, e_{s,j-1}|\{\gamma\}_1^{j-1}, \{e_s\}_1^{j-2})}. \quad (86)$$

$\gamma_j$ is iid before the attack. Therefore, applying the chain rule to $f_{a,j}(\cdot|\cdot), \quad (86)$ can be written as given in (33). Using a similar argument, (34) can be derived.

To derive (35), (38) and (40), we rewrite (30) for the case, $k > \Gamma$, applying (11) as

$$\gamma_j = w_{a,j-1} + A_a z_{j-1} - C (A + BL) \hat{x}_{j-1|j-1} - C B e_{s,j-1}. \quad (87)$$

Using the recursive state equations and the Kalman time update equation, (87) can be written as

$$\gamma_j = w_{a,j-1} + (A_a C - C (A + BL)) \left( (A + BL)^{j-2} x_{1|1} \right) + \sum_{r=1}^{j-2} (A + BL)^{r-1} B e_{s,r-1} + \sum_{r=2}^{j-2} (A + BL)^{r-1} K \gamma_{j-r} \right) - C B e_{s,j-1} + (A_a - C (A + BL) K) \gamma_{j-1}. \quad (88)$$

From the assumptions that $(A + BL)$ is strictly stable and the system started at $j = -\infty$, we can say $(A + BL)^{j-2} \to 0$ as $j \to \infty$. Therefore, (88) will take the following form as $j \to \infty$,

$$\gamma_j = w_{a,j-1} + (A_a C - C (A + BL)) \left( \sum_{r=1}^{j-2} (A + BL)^{r-1} B e_{s,j-r-1} + \sum_{r=2}^{j-2} (A + BL)^{r-1} K \gamma_{j-r} \right) - C B e_{s,j-1} + (A_a - C (A + BL) K) \gamma_{j-1}. \quad (89)$$

Using (87) and (89), we can derive the following,

$$\mu_{1,j} = E_1 \left[ \gamma_j | \{\gamma\}_1^{j-1}, \{e_s\}_1^{j-1} \right] = E_1 \left[ \gamma_j | z_{j-1}, \hat{x}_{j-1|j-1}, e_{s,j-1} \right] = A_a z_{j-1} - C (A + BL) \hat{x}_{j-1|j-1} - C B e_{s,j-1}, \quad (90)$$

$$\Sigma_{1,j} = \text{cov} \left( \gamma_j | z_{j-1}, \hat{x}_{j-1|j-1}, e_{s,j-1} \right) = Q_a. \quad (91)$$

Using the same approach, (36), (39) and (41) can be derived from (30) for the case, $k = \Gamma$. Taking expectations on both sides of (29), we get $E_0[\gamma_j] = 0$. (42) is derived from (6) as,

$$\gamma_j = y_j - C \hat{x}_{j|j-1} = C (x_j - \hat{x}_{j|j-1}) + v_j, \quad \text{and}$$

$$\Sigma_0 = E_0 \left[ \gamma_j \gamma_j^T \right] = CPC^T + R. \quad (92)$$

**APPENDIX B**

**PROOF OF LEMMA 4**

The following form of the $LSR_k$ is derived by taking logarithms on both sides of (45), and combining both the conditions in (45) using an indicator function.

$$LSR_k = \lambda_k + k |\log(1 - \rho)|$$

$$+ \log \left( LSR_0 + \sum_{i=1}^{k} (1 - \rho)^{i-1} \mathcal{L}_{d,i} \exp(-\lambda_i) \mathcal{I}_{\{LSR_i < Th_s\}} \right)$$

$$+ \sum_{i=1}^{k} (1 - \rho)^{i-1} \mathcal{L}_{b,i} \exp(-\lambda_i) \mathcal{I}_{\{LSR_i \geq Th_s\}} \right), \quad (93)$$

where

$$\lambda_k = \sum_{i=1}^{k} \log (\mathcal{L}_{a,i}) \mathcal{I}_{\{LSR_i \geq Th_s\}} + \sum_{i=1}^{k} \log (\mathcal{L}_{c,i}) \mathcal{I}_{\{LSR_i < Th_s\}}. \quad (94)$$

The threshold $Th_s$ on $LSR_k$ is the same as the threshold $Th_s$ on $p_k$. $Th_s$ is derived directly from (43) as given in (58). We rewrite $\lambda_k$ by adding and subtracting $\sum_{i=1}^{k} \log (\mathcal{L}_{a,i}) \mathcal{I}_{\{LSR_i < Th_s\}}$ to the right hand side of (94) as follows,

$$\lambda_k = Z_k + \sum_{i=1}^{k} \log (\mathcal{L}_{c,i}) \mathcal{I}_{\{LSR_i < Th_s\}}$$

$$- \sum_{i=1}^{k} \log (\mathcal{L}_{a,i}) \mathcal{I}_{\{LSR_i < Th_s\}}, \quad (95)$$

where $Z_k$ is given in (56). Replacing the first $\lambda_k$ in (93) by (95) and dividing the terms in $S_k$ and $l_k$, we get (53). The proof that $l_k$ is slowly changing variable is provided as follows.

The variable $l_k$ will be called slowly changing provided the following two conditions are satisfied, according to (46).

$$C1: k^{-1} \max \{|l_1|, \cdots, |l_k|\} \to 0, k \to \infty \quad (96)$$
and for every \( \epsilon > 0 \), there exists \( k^* \) and \( \delta > 0 \), such that for all \( k \geq k^* \)

\[
P \left\{ \max_{1 \leq i \leq k \delta} | l_{k+i} - l_k | > \epsilon \right\} < \epsilon. \tag{97}
\]

Let \( k \) from (55) be represented as the summation of three terms as follows,

\[
l_k = l_{1,k} + l_{2,k} - l_{3,k}, \tag{98}
\]

where

\[
l_{1,k} = \log \left( LSR_0 + \sum_{i=1}^{k} (1 - \rho)^{i-1} L_{d,i} \exp(-\lambda_i) \mathbb{1}_{\{LSR_i < T_h^S}\} \right)
+ \sum_{i=1}^{k} (1 - \rho)^{i-1} L_{b,i} \exp(-\lambda_i) \mathbb{1}_{\{LSR_i \geq T_h^S\}} \right), \tag{99}
\]

\[
l_{2,k} = \sum_{i=1}^{k} \log (L_{c,i}) \mathbb{1}_{\{LSR_i < T_h^S\}}, \quad \text{and} \quad \tag{100}
\]

\[
l_{3,k} = \sum_{i=1}^{k} \log (L_{a,i}) \mathbb{1}_{\{LSR_i < T_h^S\}}. \tag{101}
\]

Taking absolute values on both sides of (98), we can write,

\[
|l_k| \leq l_{1,k} + |l_{2,k}| + |l_{3,k}|. \tag{102}
\]

After the attack start point, \( LSR_k \) will gradually increase on average (from condition C6 in Theorem 1), and it will first cross \( T_h^S \) and then \( T_h^D \) as \( k \to \infty \). \( LSR_k \) will remain below \( T_h^S \) for a relatively short period of time compared to the time it takes to cross \( T_h^D \), since \( T_h^D \to \infty \). Therefore, \( l_{2,k} \) and \( l_{3,k} \) will converge to some finite values, say \( L_2 \) and \( L_3 \), respectively, as \( k \to \infty \). Also, \( \exp(-\lambda_k) \to 0 \) since \( \lambda_k \to \infty \) as \( k \to \infty \) from condition C6. Therefore, \( l_{1,k} \) will also converge to a finite value, say \( L_1 \), as \( k \to \infty \). Now, from (102), we can say \( l_k \) will also converge to a finite value \( l \), i.e., \( l \leq L_1 + L_2 + L_3 \) as \( k \to \infty \), which means \( l_k \) will satisfy condition C1.

We assume that at \( k = k_1 \), \( LSR_k \) crosses \( T_h^S \). Therefore, for \( k \geq k_1 \), we can write

\[
l_{2,k+i} = l_{2,k} = \sum_{i=1}^{k_1} \log (L_{c,i}) \mathbb{1}_{\{LSR_i < T_h^S\}}, \tag{103}
\]

\[
l_{3,k+i} = l_{3,k} = \sum_{i=1}^{k_1} \log (L_{a,i}) \mathbb{1}_{\{LSR_i < T_h^S\}}, \tag{104}
\]

\[
l_{1,k+i} = \log \left( LSR_0 + \sum_{i=1}^{k_1} (1 - \rho)^{i-1} L_{d,i} \exp(-\lambda_i) \mathbb{1}_{\{LSR_i < T_h^S\}} \right)
+ \sum_{j=1}^{k_1} (1 - \rho)^{j-1} L_{b,j} \exp(-\lambda_j) \mathbb{1}_{\{LSR_j \geq T_h^S\}} \right), \tag{105}
\]

\[
l_{1,k} = \log \left( LSR_0 + \sum_{i=1}^{k_1} (1 - \rho)^{i-1} L_{d,i} \exp(-\lambda_i) \mathbb{1}_{\{LSR_i < T_h^S\}} \right)
+ \sum_{j=1}^{k_1} (1 - \rho)^{j-1} L_{b,j} \exp(-\lambda_j) \mathbb{1}_{\{LSR_j \geq T_h^S\}} \right). \tag{106}
\]

Therefore, \( l_{k+i} - l_k = l_{1,k+i} - l_{1,k} \) for \( k \geq k_1 \).

As mentioned before, \( \exp(-\lambda_k) \to 0 \) as \( k \to \infty \), therefore, we can say \( P \{ l_{k+i} - l_k \} \to 0 \) for a sufficiently large \( k \), say \( k^* \), and \( k^* \geq k_1 \). From (107), for \( k \geq k^* \), \( P \{ |l_{k+i} - l_k| > \epsilon \} = 0 \), which in turn will satisfy condition C2.

**APPENDIX C**

**PROOF OF THEOREM 1**

First, we will show that the conditions C1-C4 are satisfied for the problem under study. \( Z_k \) is a function of continuous random variables, which take uncountably infinite values, so \( Z_k \) is non-arithmetic, thus satisfies the condition C1.

For condition C2, \( Z_1 \) denotes the log-likelihood ratio just after the attack start point. For simplicity, we consider that the attacker is present in the system from the beginning. Now, from (56) and (33), we can write \( Z_1 \) as

\[
Z_1 = -\frac{1}{2} \log \frac{|\Sigma_{1,1}|}{|\Sigma_0|} - \frac{1}{2} (\tilde{\gamma}_1 - \mu_{1,1})^T \Sigma_{1,1}^{-1} (\tilde{\gamma}_1 - \mu_{1,1})
+ \frac{1}{2} \tilde{\gamma}_1^T \Sigma_{1,1}^{-1} \tilde{\gamma}_1. \tag{108}
\]

From (38), (40), and (42), we can say that all the elements of (108) are either finite or having Gaussian distributions with finite means and variances, which in turn ensures that \( E_1 [Z_1] \) is finite.

Condition C3 is proven in Appendix D. From the expressions of \( \Sigma_0 \) (42) and \( \Sigma_\gamma \) (67), we can say that under the practical assumptions of the plant model and attacker’s system parameters \( 0 < E_1 [D (f^*, f_0)] < \infty \) from (65). In a similar way we can also show that \( 0 < E_1 [D (f_0, f_1^*)] < \infty \). Therefore, the condition C4 is valid for the problem under study. The following is the proof of Theorem 1.

Say, after the attack start point, at \( k = \bar{\lambda} \) the test statistics \( LSR_k \) will cross the threshold \( T_h^D \) at \( k = \lambda \) for the first time which is equivalent to the test statistics \( p_k \) crossing the threshold \( T_d \). To derive the expression of ADD, we assume, \( T_d = \tau - \Gamma \). After adding and subtracting \( T_d \) to (55) and rearranging the terms, it will take the following form at \( k = T_d \).

\[
S_{T_d} = T_d + (LSR_{T_d} - T_d) - l_{T_d}. \tag{109}
\]

According to the nonlinear renewal theory (46), the overshoot statistics of \( LSR_{T_d} - T_d \) can be approximated by the overshoot statistics of \( S_{T_d} \), i.e., \( r_{n_d} = S_{T_d} - T_d \), provided \( T_d \to \infty \). Moreover, the slowly changing term \( l_k \to l \) as \( k \to \infty \), where \( l \) is a RV [31]. Taking expectations on both sides of (109), we get

\[
E_1 [S_{T_d}] = T_d + \bar{\lambda} - \bar{l} + o(1), \tag{110}
\]

where \( \bar{\lambda} = \lim_{n_d \to \infty} E_1 [r_{n_d}] \) and \( \bar{l} = \lim_{k \to \infty} E_1 [l_k] \). The following expression of \( E_1 [S_{T_d}] \) is derived by taking expectations on both sides of (54) [31],

\[
E_1 [S_{T_d}] = E_1 [T_D] (E_1 [Z_1] + | \log (1 - \rho) |). \tag{111}
\]

Furthermore, \( E_1 [Z_1] \) can be approximated as \( E_1 [D (f^*, f_0)] \) as explained in [39]. Combining (110) and (111), and rearranging the term we get (60), where \( ADD = E_1 [T_D] \).
A brief derivation of FAR is provided as follows. A detailed one can be found in [31].

\[ FAR = E^x [1 - p_x] = E^x \frac{1}{1 + \rho \exp (- LSR_x)} \]  \[ \text{using (43)} \]  \[ = E^x \frac{1}{\exp (- LSR_x) + 1} \]  \[ \text{using (12)} \]

False alarm will occur when \( LSR_k \) crosses \( Th^D \) during the normal system operation. Therefore, \( \exp (- Th^D) \leq \exp (- LSR_x) \) \( \to 0 \) as \( Th^D \to \infty \), So, (12) can be approximated as

\[ FAR = \frac{1}{\rho} E^x \left[ \frac{1}{\exp (- LSR_x)} \right] (1 + o(1)) \]

where \( \rho = \exp (- Th^D) \). Replacing \( E^x \left[ \exp (- Th^D - LSR_x) \right] \) can be approximated by \( \xi \) using the overshoot \( r_{na} \) statistics [46] as given in [64]. Replacing \( E^x \left[ \exp (- Th^D - LSR_x) \right] \) by \( \xi \) in (113), we get (61).

### APPENDIX D

**PROOF OF THEOREM 2**

The proposed parsimonious watermarking mechanism can be assumed to be a always present watermarking scheme, where the watermarking signal is \( e_{s,k} = s_{k-1}e_k \). The \( e_k \) are assumed to be uncorrelated since they are generated from two independent processes. The variance of \( e_{s,k} \), \( \Sigma_{e_s} \), is derived as

\[ \Sigma_{e_s} = E_0 \left[ s_k^2 \right] = E_0 \left[ s_k^2 \right] \]  \[ \text{where (14)} \]  \[ = E_0 \left[ s_k^2 e_k^2 \right] \]

Since \( s_k = 1 \) or \( 0 \), \( E_0 \left[ s_k^2 \right] = E_0 \left[ s_k \right] \). Therefore, \( \Sigma_{e_s} \) takes the following form

\[ \Sigma_{e_s} = E_0 \left[ s_k \right] \Sigma_e = ANW \frac{E[\Gamma]}{E} \]  \[ \text{substituting (15)} \]  \[ = \rho ANW \Sigma_e \text{ since } \Gamma \sim Geom (\rho) \]

Now, the increase in the control cost, \( \Delta LQG \), is derived using Theorem 3 from [9] for the always present watermarking signal \( e_{s,k} \) as follows,

\[ \Delta LQG = tr(\Sigma_{e_s}) = \rho ANW tr(\Sigma_{e_s}) \]

where \( \Sigma_{e_s} \) is given in [73].

### APPENDIX E

#### SYSTEM PARAMETERS

The following system parameters are used for simulation study, \( \rho = 0.001 \).

**System-A parameters:**

\[ A = \begin{bmatrix} 0.75 & 0.2 \\ 0.2 & 1.0 \end{bmatrix}, \quad B = \begin{bmatrix} 0.9 & 0.5 \\ 0.1 & 1.2 \end{bmatrix}, \quad C = \begin{bmatrix} 1.0 & -1.0 \end{bmatrix} \]

**System-B parameters:**

\[ A = \begin{bmatrix} 0.968 & 0 & 0.082 & 0 \\ 0 & 0.978 & 0 & 0.064 \\ 0 & 0 & 0.917 & 0 \\ 0 & 0 & 0 & 0.935 \end{bmatrix}, \quad B = \begin{bmatrix} 0.164 & 0.004 \\ 0.002 & 0.124 \\ 0 & 0.092 \\ 0.060 & 0 \end{bmatrix} \]

**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. Alguliyev, Y. Imamverdiyev, and L. Sukhostat, “Cyber-physical systems and their security issues,” Comput. Ind., vol. 100, no. July 2017, pp. 212–223, 2018. [Online]. Available: https://doi.org/10.1016/j.compind.2018.04.017

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

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

[5] M. Abrams and J. Weiss, “Malicious control system cyber security attack case study – Maroochy Water Services,” vol. no. 1, 2019.

[6] A. Cardenas, S. Amin, B. Sinopoli, A. Gian, A. Perrig, S. Sastry et al., “Challenges for securing cyber physical systems,” in Workshop on future directions in cyber-physical systems security, vol. 5, no. 1, 2009.

[7] S. Salimi, S. Dey, and A. Ahlen, “Sequential detection of deception attacks in networked control systems with watermarking,” 2019 18th European Control Conference, ECC 2019, pp. 883–890, 2019.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] E. Mousavinejad, F. 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.

[15] 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, p. 108557, sep 2019.

[16] W. H. Ko, B. Satchidanandan, and P. R. Kumar, “Dynamic watermarking-based defense of transportation cyber-physical systems,” ACM Transactions on Cyber-Physical Systems, vol. 4, no. 1, 2019.

[17] 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.

[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] Y. Mo and B. Sinopoli, “Secure control against replay attacks,” 2009 47th Annul Allert. Conf. Commun. Control. Comput. Allert. 2009, pp. 911–918, sept 2009.

[21] A. Naha, A. M. H. Teixeira, A. Ahlen, and S. Dey, “Sequential detection of replay attacks,” IEEE Transactions on Automatic Control, pp. 1–1, 2022.

[22] C. Trapiello, D. Rotondo, H. Sanchez, and V. Puig, “Detection of replay attacks in CPS using observer-based signature compensation,” 2019 6th Int. Conf. Control. Decis. Inf. Technol. CoDIT 2019, pp. 1–6, 2019.
18

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

[24] D. Ye, T. Y. Zhang, and G. Guo, “Stochastic coding detection scheme in cyber-physical systems against replay attack,” Inf. Sci. (Ny.), vol. 481, no. 6173097, pp. 432–444, 2019. [Online]. Available: [https://doi.org/10.1016/j.ins.2018.12.091]

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

[26] A. Tartakovsky, I. Nikiforov, and M. Basseville, Sequential analysis: Hypothesis testing and changepoint detection. CRC Press, 2014.

[27] G. Lorden et al., “Procedures for reacting to a change in distribution,” The Annals of Mathematical Statistics, vol. 42, no. 6, pp. 1897–1908, 1971.

[28] A. G. Tartakovsky, “On optimum methods in quickest detection problems,” Theory Probab. Its Appl., vol. 8, no. 1, pp. 22–46, 1963.

[29] C. D. Fuh and A. G. Tartakovsky, “Asymptotic Bayesian theory of quickest change detection for hidden markov models,” IEEE Transactions on Information Theory, vol. 65, no. 1, pp. 511–529, 2019.

[30] B. Yakir, “Optimal detection of a change in distribution when the observations form a Markov chain with a finite state space,” Lecture Notes-Monograph Series, vol. 23, pp. 346–358, 1994.

[31] A. G. Tartakovsky and V. V. Veeravalli, “General asymptotic Bayesian theory of quickest change detection,” Theory of Probability and its Applications, vol. 49, no. 3, pp. 458–497, 2005.

[32] A. N. Shiryaev, “On asymptotic optimality in sequential changepoint detection: Non-iid case,” IEEE Trans. Inf. Theory, vol. 63, no. 6, pp. 3433–3450, 2017.

[33] K. Premkumar and A. Kumar, “Optimal sleep-wake scheduling for quickest intrusion detection using sensor networks,” no. ii, pp. 2074–2082, 2008.

[34] T. Banerjee and V. V. Veeravalli, “Data-Efficient Quick Change Detection with On-Off Observation Control,” Seq. Anal., vol. 31, no. 1, pp. 40–77, 2012.

[35] A. Naha, A. Teixeira, A. Ahlen, and S. Dey, “Deception attack detection using reduced watermarking,” 2021 European Control Conference (ECC), pp. 74–80, 2021.

[36] M. H. Yılmaz and 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.

[37] Y.-C. Liu, G. Bianchin, and F. Pasqualetti, “Secure trajectory planning against undetectable spoofing attacks,” Automatica, vol. 112, p. 108655, 2020.

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

[39] A. Naha, A. Teixeira, A. Ahlén, and S. Dey, “Quickest detection of deception attacks in networked control systems with physical watermarking,” arXiv preprint arXiv:2101.01466, 2021.

[40] ———, “Sequential detection of replay attacks with a parsimonious watermarking policy,” accepted in American Control Conference, 2022.

[41] E. Jonsson and T. Olovsson, “A quantitative model of the security intrusion process based on attacker behavior,” IEEE Transactions on Software Engineering, vol. 23, no. 4, pp. 235–245, 1997.

[42] F. Arnold, H. Hermanns, R. Pulungan, and M. Stoelinga, “Time-dependent analysis of attacks,” in International Conference on Principles of Security and Trust. Springer, 2014, pp. 285–305.

[43] B. Prochaska, “A note on the relationship between the geometric and exponential distributions,” The American Statistician, vol. 27, no. 1, pp. 27–27, 1973.

[44] F. J. Beutler and K. W. Ross, “Optimal policies for controlled markov chains with a constraint,” Journal of Mathematical Analysis and Applications, vol. 112, no. 1, pp. 236–252, 1985.

[45] D. P. Bertsekas, Dynamic programming and optimal control. Athena scientific Belmont, MA, 1995, vol. 1, no. 2.

[46] D. Siegmund, Sequential analysis: tests and confidence intervals. Springer Science & Business Media, 2013.

[47] H. Akaike, “Fitting autoregressive models for prediction,” Annals of the institute of Statistical Mathematics, vol. 21, no. 1, pp. 243–247, 1969.

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

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