Quickest detection of deception attacks on cyber–physical systems with a parsimonious watermarking policy☆

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

[a] Electrical Engineering, Uppsala University, Uppsala, Sweden
[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.

1. 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. Satchidanandan and Kumar (2017). 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 (Alguliyev, Imamverdiyev, & Sukhostat, 2018; Satchidanandan & Kumar, 2017). 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 (Mo, Weerakkody, & Sinopoli, 2015). 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 (Langner, 2011), 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 (Abrams & Weiss, 2008), the attack on the Davis–Besse nuclear power plant in Ohio, USA (Cardenas et al., 2009), 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 (Satchidanandan & Kumar, 2017).

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 (Mo et al., 2015; Satchidanandan & Kumar, 2017). Replay attacks are one kind of data deception

attack, where the attacker replaces the true observations with previously recorded data to remain stealthy (Mo et al., 2015). In DoS attacks, the attacker's objective is to disrupt the availability of data. The attacker may achieve that by overpowering the wireless network (Salimi, Dey, & Ahlen, 2019). 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.

1.1. 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 (Chen, Kar, & Moura, 2018; Park, Lee, Shim, Eun, & Johansson, 2019), attack resilient state estimation (Du et al., 2019; Fawzi, Tabuada, & Diggavi, 2014; Forti et al., 2018), attack detection strategies (Fang, Qi, Cheng, & Zheng, 2020; Ge, Han, Zhong, & Zhang, 2019; Ko, Satchidanandan, & Kumar, 2019; Mo, Chabukswar, & Sinopoli, 2014; Mo et al., 2015; Mousavinejad, Yang, Han, & Vlacic, 2018; Pasqualetti, Dorfler, & Bullo, 2013; Satchidanandan & Kumar, 2017, 2020), 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 (Ge et al., 2019; Mousavinejad et al., 2018; Pasqualetti et al., 2013). However, as studied in the literature, passive detection schemes generally have an unsatisfactory probability of detection in the presence of noise and uncertainties (Mo & Sinopoli, 2009). 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 Mo and Sinopoli (2009) to detect replay attacks by adding an iid watermarking signal to the control input and performing a χ^2 test using the innovation signal from the state estimator. The method in Mo and Sinopoli (2009) is improved by designing an optimal watermarking signal in Mo et al. (2014). Instead of an iid watermarking scheme, the watermarking signal generated from a hidden Markov Model (HMM) is studied in Mo et al. (2015). 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 Naha, Teixeira, Ahlen, and Dey (2022b). Besides the innovation signal, the observation signal is also used to generate residue signals for the attack detections (Satchidanandan & Kumar, 2017). In Mo et al. (2015), Naha et al. (2022b) and Satchidanandan and Kumar (2017), 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 (Mo et al., 2015). The relation between the increase in the linear quadratic Gaussian control cost, ΔLQG, and the watermarking signal variance is studied in Mo et al. (2015). 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 (Fang et al., 2020) 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 (Fang et al., 2020). Another approach is to add watermarking directly to the observations (Ferrari & Teixeira, 2017; Trapiello, Rotondo, Sanchez, & Puig, 2019; Ye, Zhang, & Guo, 2019). 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 (Ferrari & Teixeira, 2017), time-varying sinusoidal (Sánchez et al., 2019), random noise (Ye et al., 2019), multiplicative to the observations (Trapiello et al., 2019), 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 (Tartakovsky, Nikiforov, & Basseville, 2014). 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 (Lorden et al., 1971). 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 (1963). The original Shiryaev rule was proposed for the data with different iid distributions, before and after the change point. Finding an optional detection rule for the general non-iid data is difficult (Fuh & Tartakovsky, 2019). In Yakir (1994), an optimal detection rule is developed for homogeneous finite-state Markov chains. A slightly different approach is followed in Tartakovsky (2017) and Tartakovsky and Veeravalli (2005), 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 Tartakovsky (2017). 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 $\to \infty$ and a few other conditions are satisfied. The condition (3) of Tartakovsky (2017) for the optimality is that the prior distribution of the change point must satisfy (1).

$$\lim_{k \to \infty} \frac{\log \mathrm{P}\{\Gamma \geq k+1\}}{k} = -c, \ c \geq 0, \tag{1}$$

where Γ 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 (Fuh & Tartakovsky, 2019). 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 (Fang et al., 2020). 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) (Tartakovsky et al., 2014). Additionally, our work in this paper is inspired by two other prior works (Banerjee & Veeravalli, 2012; Premkumar & Kumar, 2008). The quickest intrusion detection problem is studied in Premkumar and Kumar (2008), 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 Banerjee and Veeravalli (2012) 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 (Banerjee & Veeravalli, 2012; Premkumar & Kumar, 2008), we have also assumed the distribution of the change point, *i.e.*, the attack start point, to be a geometric distribution with parameter ρ, which satisfies the condition given in (1).

1.2. Contributions

In our previous work (Naha et al., 2022b), we studied in detail that the worst-case ADD decreases with the increase in ΔLQG, which denotes the increases in the LQG control cost due to the addition of watermarking for the always-present watermarking scheme. Additionally, ΔLQG 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 ΔLQG. 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 (Tartakovsky & Veeravalli, 2005). 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 Th^s and Th^d, $Th^d \geq Th^s$, on the posterior probability of attack p_k under practical circumstances. In other words, if $p_k \geq Th^s$, then we add watermarking to the $(k+1)$th control input. On the other hand, if $p_k \geq Th^d$, we decide that the attack is present in the system and terminate the process. Our study shows that Th^s primarily controls the ANW, which in turn controls the ΔLQG value, and Th^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 ΔLQG. We have also obtained an expression of ΔLQG for a given ANW. We have reported a preliminary simulation study on this problem previously for a single-input single-output (SISO) system in Naha, Teixeira, Ahlen, and Dey (2021a). In the current

Table 1
Notations.

Symbol	Description
$\mathbb{R}^n$	The set of $n \times 1$ real vectors
$\mathbb{R}^{m \times n}$	The set of $m \times n$ real matrices
$\hat{(\cdot)}$	Estimated quantity
$E[\cdot]$	Expectation operator
$\hat{\mathbf{x}}_{k|k}$	Estimated state at kth instant using measurements up to kth instant
$[\cdot]^T$	Transpose of a matrix or vector
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and variance Σ
$\{\cdot\} \cup \{\cdot\}$	Union of two sets
$\Sigma \geq 0$	Σ is a positive semi-definite matrix
$\Sigma > 0$	Σ is a positive definite matrix
$\mathbf{x}_{a,k}, \mathbf{u}_{d,k},$ $\mathbf{e}_{s,k}$, etc.	kth instant values of $\mathbf{x}_a$, $\mathbf{u}_d$, $\mathbf{e}_s$, etc.
$\{\cdot\}^*$	Optimum quantity
$[\cdot]_{ij}$	ith row and jth column element of a matrix
$P\{\cdot\}$	Probability measure
Π_k	Probability of the event $\{\Gamma = k\}$
P_k	Probability measure when the change point $\Gamma = k$
$P^\Pi\{\cdot\}$	Average probability measure, $= \sum_{k=1}^{\infty} \Pi_k P_k\{\cdot\}$
E^Π	Expectation with respect to probability measure P^Π
$f_{1,j}, f_{2,j}, \mu_{1,j},$ $\mu_{2,j}, \Sigma_{1,j}, \Sigma_{2,j}$	jth instant values of f_1, f_2, μ_1, μ_2, Σ_1, Σ_2
$\|\cdot\|$	Determinant of a matrix or absolute value of a scalar
$\bar{\{\cdot\}}$	Mean value of a quantity
$\text{tr}(\cdot)$	Trace of a matrix
$\{\mathbf{X}\}_1^{k-1}$	$\{X_i : 1 \leq i \leq k-1\}$
$\mathbb{1}_{\{condition\}}$	Indicator function, 1 if condition is true, 0 otherwise

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 ΔLQG that facilitate the optimal design of the watermarking process.
(3) We have optimized the watermarking signal variance to maximize KLD, which improves ADD for a fixed upper bound on the ΔLQG.

The paper is organized as follows. Section 2 discusses the system model and the attack strategy considered in this paper. The defence mechanism is explained in Section 3. Section 4 provides the analytical expressions of ADD, FAR and the relationship between the ANW and ΔLQG. It also explains the optimization framework for the watermarking signal variance. Section 5 presents and discusses the simulation results. Section 6 concludes the paper.

1.3. Notations

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

2. System model

The system model during normal operations and the model with the data deception attack are discussed in this section.
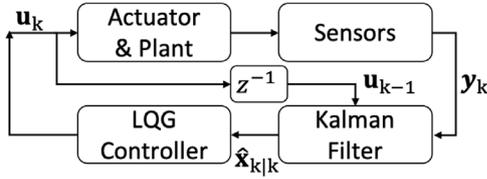
Fig. 1. Schematic diagram of the system during normal operation.

Fig. 2. Schematic diagram of the system under attack.

2.1. 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 MIMO plant with the following state update and measurement equations,

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k + \mathbf{w}_k, \tag{2}$$

$$\mathbf{y}_k = \mathbf{C}\mathbf{x}_k + \mathbf{v}_k, \tag{3}$$

where $\mathbf{x}_k \in \mathbb{R}^n$ and $\mathbf{y}_k \in \mathbb{R}^m$ are the state and measurement vectors, respectively. $\mathbf{u}_k \in \mathbb{R}^p$ is the control input vector. The process and observation noise vectors are denoted as $\mathbf{w}_k \in \mathbb{R}^n \sim \mathcal{N}(0, \mathbf{Q})$ and $\mathbf{v}_k \in \mathbb{R}^m \sim \mathcal{N}(0, \mathbf{R})$, respectively, with $\mathbf{Q} > \mathbf{0}$ and $\mathbf{R} > \mathbf{0}$. Here, $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{n \times p}$, $\mathbf{Q} \in \mathbb{R}^{n \times n}$, $\mathbf{C} \in \mathbb{R}^{m \times n}$, and $\mathbf{R} \in \mathbb{R}^{m \times m}$. 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 are as follows,

$$\hat{\mathbf{x}}_{k|k-1} = \mathbf{A}\hat{\mathbf{x}}_{k-1|k-1} + \mathbf{B}\mathbf{u}_{k-1}, \tag{4}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}\gamma_k, \tag{5}$$

where $\hat{\mathbf{x}}_{k|k-1} = \mathrm{E}[\mathbf{x}_k|\mathcal{I}_{k-1}]$ and $\hat{\mathbf{x}}_{k|k} = \mathrm{E}[\mathbf{x}_k|\mathcal{I}_k]$ are the Kalman predicted and filtered states, respectively. $\mathrm{E}[\cdot]$ denotes the expectation operator, and $\mathcal{I}_k \triangleq \{\mathbf{u}_0, \mathbf{u}_1, \ldots \mathbf{u}_k, \mathbf{y}_0, \mathbf{y}_1, \ldots \mathbf{y}_k\}$ is the set of all information up to time k. The innovation signal γ_k and the steady state Kalman filter gain $\mathbf{K}$ are given as,

$$\gamma_k = \mathbf{y}_k - \mathbf{C}\hat{\mathbf{x}}_{k|k-1}, \tag{6}$$

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

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

$$\mathbf{P} = \mathbf{A}\mathbf{P}\mathbf{A}^T + \mathbf{Q} - \mathbf{A}\mathbf{P}\mathbf{C}^T \left(\mathbf{C}\mathbf{P}\mathbf{C}^T + \mathbf{R}\right)^{-1} \mathbf{C}\mathbf{P}\mathbf{A}^T. \tag{8}$$

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 $\mathbf{u}_k^*$ is derived by minimizing the following cost function,

$$J = \lim_{T \to \infty} \mathrm{E}\left[\frac{1}{2T+1}\left\{\sum_{k=-T}^{T} \left(\mathbf{x}_k^T \mathbf{W}\mathbf{x}_k + \mathbf{u}_k^T \mathbf{U}\mathbf{u}_k\right)\right\}\right] \tag{9}$$

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

$$\mathbf{L} = -\left(\mathbf{B}^T \mathbf{S}\mathbf{B} + \mathbf{U}\right)^{-1} \mathbf{B}^T \mathbf{S}\mathbf{A}.$$
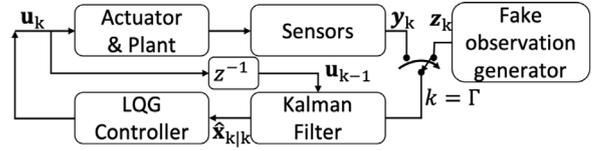
Here, $\mathbf{S}$ is the solution to the following algebraic Riccati equation,

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

2.2. 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 $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{Q}$, $\mathbf{R}$, and $\mathbf{L}$;
(3) the attacker cannot access or alter the control signal.

To launch a data deception attack, the attacker replaces the true observations $\mathbf{y}_k$ by the fake data $\mathbf{z}_k$ 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 (Liu, Bianchin, & Pasqualetti, 2020; Yılmaz & Arslan, 2015). The fake observation data $\mathbf{z}_k$ is assumed to be generated from the following linear stationary stochastic process,

$$\mathbf{z}_k = \mathbf{A}_a \mathbf{z}_{k-1} + \mathbf{w}_{a,k-1}, \tag{11}$$

where $\mathbf{w}_{a,k} \sim \mathcal{N}(0, \mathbf{Q}_a)$ is the iid noise vector at the kth time instant, and $\mathbf{Q}_a \in \mathbb{R}^{m \times m}$. A similar attack strategy is also studied in Li and Ye (2022), 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.

Remark 1. In general, for linear control systems, the measurement vector $\mathbf{y}_k$ 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 (11). In other words, the attack model in (11) mimics the linear stationary stochastic model of $\mathbf{y}_k$, which makes it challenging to detect. In addition to that, the attack model in (11) can make closed-loop control systems unstable, as discussed in Naha, Teixeira, Ahlén, and Dey (2021b), 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 (11) 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 Naha, Teixeira, Ahlén, and Dey (2022c).

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

$$\gamma_k = \mathbf{z}_k - \mathbf{C}\hat{\mathbf{x}}_{k|k-1}, \tag{12}$$

whereas $\hat{\mathbf{x}}_{k|k-1}$ and $\hat{\mathbf{x}}_{k|k}$ will follow the same time update and measurement update equations (4) and (5), respectively. However, the γ_k in (5) for $k \geq \Gamma$ will follow (12). 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 $\mathbf{y}_k$ by the fake data $\mathbf{z}_k$ from $k \geq \Gamma$, the innovation signal γ_k automatically takes the form given in (12).

As studied in several works of literature, the distribution of the attack start point can be modelled as exponential distribution for continuous-time systems (Arnold, Hermanns, Pulungan, & Stoelinga, 2014; Jonsson & Olovsson, 1997). In Jonsson and Olovsson (1997), 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 Arnold et al. (2014). Additionally, the authors studied various practical systems, including the famous Stuxnet attack (Langner, 2011), 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 (Prochaska, 1973), we have modelled the attack start point Γ to be an RV that follows a geometric distribution with parameter ρ, where $0 < \rho < 1$. Here ρ is a design parameter reflecting the defender's belief of how often attacks occur. For the proposed method, ρ is a parameter that needs to be set by the defender. A vulnerability analysis of the system can decide the value of ρ, see Arnold et al. (2014) and Jonsson and Olovsson (1997). From the derived approximate analytical expressions of ADD (60), FAR (61) and ΔLQG (72), we can say that ADD and ΔLQG will not be affected much by the difference in the chosen ρ and the attacker's true ρ, 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 ρ is higher than the attacker's true ρ and vice versa. The defender can thus choose a suitable ρ, depending on the specification on the maximum false alarm rate. Finally, we can write the prior probability $\Pi_k \triangleq \mathrm{P}\{\Gamma = k\}$ in the following form (Banerjee & Veeravalli, 2012),

$$\Pi_k = \Pi_0 \mathbb{1}_{\{k=0\}} + (1 - \Pi_0) \rho (1 - \rho)^{k-1} \mathbb{1}_{\{k \geq 1\}}. \tag{13}$$

Here, $\Pi_0 \triangleq \mathrm{P}\{\Gamma \leq 0\}$, *i.e.*, Π_0 is the prior probability of the attack happening before the start of the observation. $\mathbb{1}_{\{condition\}}$ 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 Γ, but knows the prior distribution of Γ.

3. 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

$$\mathbf{e}_k \sim \mathcal{N}(\mathbf{0}, \Sigma_e), \tag{14}$$

to the optimal LQG control input, $\mathbf{u}_k^*$, 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 kth time
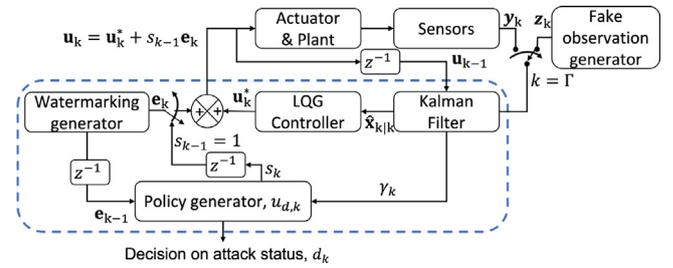
Fig. 3. Schematic diagram of the system with the proposed watermarking scheme.

instant is controlled by the optimal policy $\mathbf{u}_d^*$. Here, the subscript d of $\mathbf{u}_d^*$ indicates that the optimal policy is derived using dynamic programming. The policy $\mathbf{u}_{d,k}$ decides the values of the following two control variables s_k and d_k at the kth time instant,

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

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

$$\mathbf{u}_k = \mathbf{u}_k^* + s_{k-1}\mathbf{e}_k. \tag{17}$$

Fig. 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.

3.1. Problem formulation

Our objective is to find the optimal policy $\mathbf{u}_d^*$ 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 Tartakovsky and Veeravalli (2005).

False alarm rate (FAR): FAR is defined as

$$FAR \triangleq \mathrm{P}^{\Pi}\{\tau < \Gamma\}. \tag{18}$$

Here, P^{Π} indicates the average probability measure. $\mathrm{P}^{\Pi}\{\Omega\} = \sum_{k=1}^{\infty} \Pi_k \mathrm{P}_k\{\Omega\}$, where Ω is any event and P_k is the probability measure when the change point $\Gamma = k$. τ is the time instant when the hypothesis H_1 is selected.

Average detection delay (ADD): ADD is defined as

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

Here, E^{Π} denotes the expectation with respect to the probability measure P^{Π}.

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

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

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

Now, we formulate the following optimization problem,

$$\min_{\mathbf{u}_d} ADD,$$
$$\text{s.t. } FAR \leq FAR_{th}, \tag{21}$$
$$ANW \leq ANW_{th},$$

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 Banerjee and Veeravalli (2012) and Beutler and Ross (1985), except the term $\lambda_e ANW$, and it reads as

$$J^* = \min_{\mathbf{u}_d} \ ADD + \lambda_f FAR + \lambda_e ANW, \tag{22}$$

where $\lambda_f \geq 0$ and $\lambda_e \geq 0$ are the Lagrangian multipliers. A new state variable θ_k is defined as

$$\theta_k \triangleq \begin{cases} 0, & \text{No attack,} \\ 1, & \text{System under attack,} \\ T_e, & \text{Attack detected by hypothesis testing} \\ & \text{and process terminated.} \end{cases} \tag{23}$$

Similar to Banerjee and Veeravalli (2012), ADD, FAR and ANW can also be expressed in terms of the control variables, s_k and d_k, and the state variable θ_k as follows,

$$ADD = \mathrm{E}\left[\sum_{k=1}^{\tau} \mathbb{1}_{\{\theta_k=1\}}\mathbb{1}_{\{d_k=0\}}\right], \tag{24}$$

$$FAR = \mathrm{E}\left[\sum_{k=1}^{\tau} \mathbb{1}_{\{\theta_k=0\}}\mathbb{1}_{\{d_k=1\}}\right], \text{ and} \tag{25}$$

$$ANW = \mathrm{E}\left[\sum_{k=1}^{\tau} \mathbb{1}_{\{\theta_k=0\}}\mathbb{1}_{\{s_k=1\}}\mathbb{1}_{\{d_k=0\}}\right]. \tag{26}$$

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

$$J^* = \min_{\mathbf{u}_d} \ \mathrm{E}\left[\sum_{k=1}^{\tau} g_k\left(\theta_k, s_k, d_k\right)\right], \tag{27}$$

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

$$\begin{aligned} g_k\left(\theta_k, s_k, d_k\right) = \mathbb{1}_{\{\theta_k \neq T_e\}} \Big[&\mathbb{1}_{\{\theta_k=1\}}\mathbb{1}_{\{d_k=0\}} \\ + \lambda_f \mathbb{1}_{\{\theta_k=0\}}\mathbb{1}_{\{d_k=1\}} &+ \lambda_e \mathbb{1}_{\{\theta_k=0\}}\mathbb{1}_{\{s_k=1\}}\mathbb{1}_{\{d_k=0\}}\Big]. \end{aligned} \tag{28}$$

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 θ_k is not observable to the defender. Therefore, we replace the state θ_k by its sufficient statistics p_k. The sufficient statistics p_k, *i.e.*, the posterior probability of attack at kth time instant is defined as, $p_k \triangleq \mathrm{P}\{\Gamma \leq k|\Psi_k\} = \mathrm{E}\left[\mathbb{1}_{\{\theta_k=1\}}|\Psi_k\right]$, where Ψ_k |. The optimization problem in (27) is then redefined and solved using p_k as discussed in details in the following Section 3.2.

The accessibility hypothesis discussed in Beutler and Ross (1985) tells us that under this hypothesis, for every stationary deterministic policy $\mathbf{u}_d \in \mathcal{U}$, any arbitrary state, say θ_k is accessible from each starting state $\theta_k = \theta_0$. Here, $\mathcal{U}$ is the set of all permissible stationary deterministic policies. Under the accessibility hypothesis, the dynamic programming equation using the cost function of (27) is solvable by at least one stationary deterministic policy for each $\lambda_f \geq 0$ and $\lambda_e \geq 0$ (Beutler & Ross, 1985).

3.2. Finding the optimal policy

This section discusses the solution approach taken to solve the optimization problem of (27) in the following three main steps.

3.2.1. Selection of test signals
Combining (2)–(6) and (12), we can represent the innovation signal as

for $k < \Gamma$,

$$\gamma_k = \mathbf{CA}\left(\mathbf{x}_{k-1} - \hat{\mathbf{x}}_{k-1|k-1}\right) + \mathbf{Cw}_{k-1} + \mathbf{v}_k, \text{ and} \tag{29}$$

for $k \geq \Gamma$,

$$\gamma_k = \mathbf{z}_k - \mathbf{C}\left(\mathbf{A} + \mathbf{BL}\right)\hat{\mathbf{x}}_{k-1|k-1} - \mathbf{CB}s_{k-2}\mathbf{e}_{k-1}. \tag{30}$$

So, the innovation signal is dependent on the watermarking signal after the attack, see (30). On the contrary, the innovation signal is independent of the watermarking signal before the attack, see (29). It is assumed that the attacker will be replacing the true stationary observation $\mathbf{y}_k$ with fake but stationary data $\mathbf{z}_k$ to remain stealthy. In addition to that, as discussed in Section 3.3, the optimal policy $\mathbf{u}_d^*$ is also a stationary one. Therefore, the innovation signal will be stationary but with different distributions before and after the attack, as $k \rightarrow \infty$. Also, from the properties of the Kalman filter, we know the innovation signal is iid before the attack. Additionally, the use of the joint statistics of the innovation signal and the watermarking signal increases the KLD compared to the case where the statistics of the innovation signal alone is used, see Theorem 1 and Remark 1 from Salimi et al. (2019), where the improvement in KLD has been quantified for a single-input single-output (SISO) system. These reasons motivate us to use the joint distribution of the innovation signal and the watermarking signal to generate the test statistics for attack detections.

3.2.2. Derivation of test statistics
We use the Shiryaev statistics because of its asymptotic optimality for a fixed upper bound on FAR as stated in Theorem 1 from Tartakovsky (2017). The data is assumed to be iid before and after the change point in Tartakovsky (2017). In contrast to Tartakovsky (2017), in our study, the innovation signal γ_k is iid before the attack and non-iid after the attack. From Tartakovsky et al. (2014), the Shiryaev statistics SR_k at the kth instant in time for our problem formulation can be written as

$$SR_k = \sum_{i=1}^{k}\prod_{j=i}^{k}\frac{\mathcal{L}_j}{1-\rho}, \tag{31}$$

where i is the candidate change or attack start point, and $\mathcal{L}_j$ is the likelihood ratio. The expression for $\mathcal{L}_j$ 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 $\mathcal{L}_j$ 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,*

$$\mathcal{L}_j = \begin{cases} \mathcal{L}_{a,j}, & j > i \\ \mathcal{L}_{b,j}, & j = i, \end{cases} \text{ and} \tag{32}$$

$$\mathcal{L}_{a,j} = \frac{f_{1,j}\left(\gamma_j|\{\gamma\}_1^{j-1}, \{\mathbf{e}_s\}_1^{j-1}\right)}{f_0\left(\gamma_j\right)}, \tag{33}$$

$$\mathcal{L}_{b,j} = \frac{f_{2,j}\left(\gamma_j|\{\gamma\}_1^{j-1}, \{\mathbf{e}_s\}_1^{j-1}\right)}{f_0\left(\gamma_j\right)}. \tag{34}$$

Here, $f_{1,j}(\cdot|\cdot)$, $f_{2,j}(\cdot|\cdot)$ and $f_0(\cdot)$ denote the distributions for $j > i$, $j = i$ and $j < i$, respectively, and $\mathbf{e}_{s,j} = s_{j-1}\mathbf{e}_j$. $f_{1,j}(\cdot|\cdot)$, $f_{2,j}(\cdot|\cdot)$ and $f_0(\cdot)$ take the following forms,

$$f_{1,j}(\cdot|\cdot) = \mathcal{N}\left(\mu_{1,j}, \Sigma_{1,j}|\mathbf{z}_{j-1}, \hat{\mathbf{x}}_{j-1|j-1}, \mathbf{e}_{s,j-1}\right), \tag{35}$$

$$f_{2,j}(\cdot|\cdot) = \mathcal{N}\left(\mu_{2,j}, \Sigma_{2,j}|\hat{\mathbf{x}}_{j-1|j-1}, \mathbf{e}_{s,j-1}\right), \tag{36}$$

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

Here

$$\mu_{1,j} = \mathbf{A}_a\mathbf{z}_{j-1} - \mathbf{C}\left(\mathbf{A} + \mathbf{BL}\right)\hat{\mathbf{x}}_{j-1|j-1} - \mathbf{CBe}_{s,j-1} \tag{38}$$

$$\mu_{2,j} = -\mathbf{C}\left(\mathbf{A} + \mathbf{BL}\right)\hat{\mathbf{x}}_{j-1|j-1} - \mathbf{CB}\mathbf{e}_{s,j-1} \tag{39}$$

$$\Sigma_{1,j} = \mathbf{Q}_a \tag{40}$$

$$\Sigma_{2,j} = \mathbf{Q}_z \tag{41}$$

$$\Sigma_0 = \mathbf{CPC}^T + \mathbf{R} \tag{42}$$

Proof. The proof of Lemma 1 can be found in Appendix A of Naha, Teixeira, Ahlén, and Dey (2022a). It is omitted from this paper due to space limitations. □

Eq. (31) is same as the original Shiryaev statistics, where Π_0 is assumed to be 0, see (6.9) from Tartakovsky et al. (2014). 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 γ_j on the previous values of γ and $\mathbf{e}_s$ from the time index 1 to $j-1$ can be approximated as given in (33)–(39), where γ_j is only dependent on the immediate past values, i.e., at time index $j-1$, of $\mathbf{z}$, $\hat{\mathbf{x}}$ and $\mathbf{e}_s$.

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

We have applied the value iteration from Bertsekas (1995) 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 Tartakovsky et al. (2014),

$$p_k = \frac{SR_k}{SR_k + 1/\rho}. \tag{43}$$

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 kth time instant, p_k, for the test data γ_k (iid (29) and non-iid (30)) and $\mathbf{e}_k$ (14), can be updated using the following recursion formula, when the attack start point is geometrically distributed with parameter ρ,*

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

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. (44)

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

$$SR_k = \begin{cases} \frac{\mathcal{L}_{c,k}}{1-\rho}SR_{k-1} + \frac{\mathcal{L}_{d,k}}{1-\rho}, & s_{k-2} = 0, \\ \frac{\mathcal{L}_{a,k}}{1-\rho}SR_{k-1} + \frac{\mathcal{L}_{b,k}}{1-\rho}, & s_{k-2} = 1. \quad\square \end{cases} \tag{45}$$

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.

Table 2

$\mathcal{U}.$

$\mathbf{u}_{d,k}$	s_k	d_k
1	0	0
2	1	0
3	0	1

3.2.3. 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 $p_k = \mathrm{E}\left[\mathbb{1}_{\{\theta_k=1\}}|\Psi_k\right]$, see (46).

$$\mathrm{E}\left[g_k\left(\theta_k, s_k, d_k\right)|\Psi_k\right] = p_k\mathbb{1}_{\{d_k=0\}} + \lambda_f\left(1-p_k\right)\mathbb{1}_{\{d_k=1\}} + \lambda_e\left(1-p_k\right)\mathbb{1}_{\{s_k=1\}}\mathbb{1}_{\{d_k=0\}}. \tag{46}$$

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

$$J(p_k) = \min_{\mathbf{u}_d \subset \mathcal{U}} \big[p_k\mathbb{1}_{\{d_k=0\}} + \lambda_f\left(1-p_k\right)\mathbb{1}_{\{d_k=1\}}$$

$$+ \lambda_e\left(1-p_k\right)\mathbb{1}_{\{s_k=1\}}\mathbb{1}_{\{d_k=0\}} + B_0(p_k)\mathbb{1}_{\{s_k=0\}}\mathbb{1}_{\{d_k=0\}}$$

$$+ B_1(p_k)\mathbb{1}_{\{s_k=1\}}\mathbb{1}_{\{d_k=0\}} \big], \tag{47}$$

where $\mathcal{U} = \{1, 2, 3\}$ is the set of all stationary deterministic permissible policies, see Table 2. $B_0(p_k) = \mathrm{E}\left[J\left(\phi_0(p_k)\right)\right]$, and $B_1(p_k) = \mathrm{E}\left[J\left(\phi_1(p_k)\right)\right]$. 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 kth 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.

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 $\mathbf{u}_d^*$.

3.3. 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 $\mathbf{u}_d$, 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.

$$s_k = \begin{cases} 0, & B_0(p_k) - B_1(p_k) < \lambda_e(1-p_k), \\ 1, & otherwise. \end{cases} \tag{48}$$

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

First, we will prove that the optimal policy is going to be a stationary deterministic policy. As stated in Lemma 3.1 from Beutler and Ross (1985), the costs FAR and ANW will be monotone and

non-increasing in λ_f and λ_e, respectively. We can prove that by following the similar steps used to prove Lemma 3.1 in Beutler and Ross (1985). 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 Beutler and Ross (1985).

Finally, as discussed in Beutler and Ross (1985), 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, λ_e and λ_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^s and Th^d, $Th^d \geq Th^s$. Figs. 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 λ_e and large λ_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 two-threshold policy. We have found that for a relatively large λ_e or ρ near to unity, the optimal policy may even become a one or three-threshold policy, which is similar to the findings of Banerjee and Veeravalli (2012). The following steps can be followed offline to find the two thresholds.

Step 1 : The search space of λ_e and λ_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 λ_e^* and λ_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 $\bar{p} = Th^s$ and $\bar{p} = 1$. Eq. (51) is derived from (49) using $s_k = 0$ and solved for $\bar{p} \in [Th^s, 1]$. The solutions of (51) is $\bar{p} = Th^d$ and $\bar{p} = 1$.

$$B_0(\bar{p}) - B_1(\bar{p}) = \lambda_e(1 - \bar{p}) \tag{50}$$

$$\bar{p} + \lambda_e(1 - \bar{p}) + B_1(\bar{p}) = \lambda_f(1 - \bar{p}) \tag{51}$$

Finally, the optimal policy $\mathbf{u}_d^*$ is given as

$$\mathbf{u}_{d,k}^* = \begin{cases} 1, \ i.e., \ (s_k = 0, d_k = 0) & p_k < Th^s, \\ 2, \ i.e., \ (s_k = 1, d_k = 0) & p_k \geq Th^s, \\ 3, \ i.e., \ (s_k = 0, d_k = 1) & p_k \geq Th^d. \end{cases} \tag{52}$$

Next, we briefly discuss the computational runtime complexity of the proposed policy.
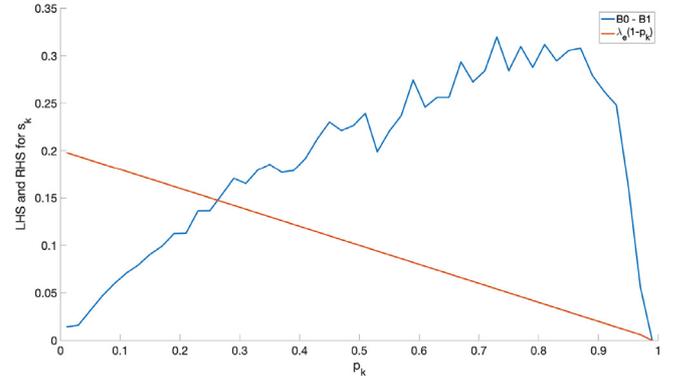
Fig. 4. LHS and RHS of (48) vs. p_k for System-A. $\lambda_e = 0.2$, $\lambda_f = 100$, and $\sigma_e^2 = 1.19$.
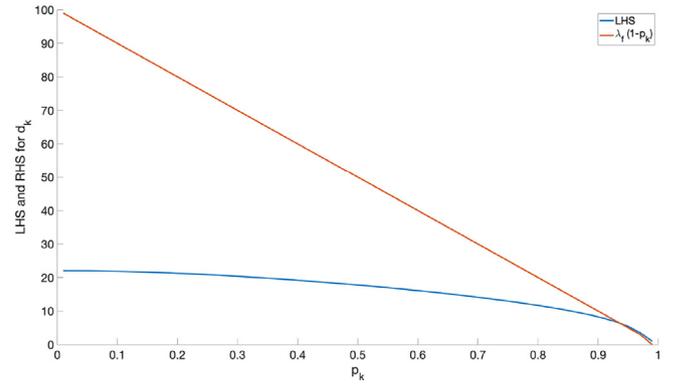
Fig. 5. LHS and RHS of (49) vs. p_k for System-A. $\lambda_e = 0.2$, $\lambda_f = 100$, and $\sigma_e^2 = 1.19$.

3.4. 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 run-time are a few matrix–vector multiplications with the highest computational complexity of $O(np)$, see (38) and (39).

4. Derivations of ADD, FAR and $\triangle LQG$

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

4.1. 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 (Siegmund, 2013; Tartakovsky & Veeravalli, 2005). 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 γ_k ((29) and (30)) and the watermarking signal $\mathbf{e}_k$ (14), under the two threshold policy Th^s*

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 (Siegmund, 2013).

$$LSR_k = S_k + l_k. \tag{53}$$

$$S_k = Z_k + k|\log(1-\rho)|. \tag{54}$$

$$l_k = \log\left(SR_0 + \sum_{i=1}^{k}(1-\rho)^{i-1}\mathcal{L}_{d,i}\exp(-\lambda_i)\mathbb{1}_{\{LSR_i < Th^S\}}\right.$$

$$\left. + \sum_{i=1}^{k}(1-\rho)^{i-1}\mathcal{L}_{b,i}\exp(-\lambda_i)\mathbb{1}_{\{LSR_i \geq Th^S\}}\right) \tag{55}$$

$$+ \sum_{j=1}^{k}\log\left(\mathcal{L}_{c,j}\right)\mathbb{1}_{\{LSR_j < Th^S\}} - \sum_{j=1}^{k}\log\left(\mathcal{L}_{a,j}\right)\mathbb{1}_{\{LSR_j < Th^S\}},$$

where

$$Z_k = \sum_{i=1}^{k}\log\left(\mathcal{L}_{a,i}\right), \tag{56}$$

$$\lambda_k = \sum_{i=1}^{k}\log\left(\mathcal{L}_{a,i}\right)\mathbb{1}_{\{LSR_i \geq Th^S\}}$$

$$+ \sum_{i=1}^{k}\log\left(\mathcal{L}_{c,i}\right)\mathbb{1}_{\{LSR_i < Th^S\}}, \text{ and} \tag{57}$$

$$Th^S = \log\frac{Th^s}{\rho\left(1-Th^s\right)}. \tag{58}$$

Proof. The proof of Lemma 3 is provided in Appendix A. □

Remark 3. The threshold Th^s for p_k is equivalent to the threshold Th^S for LSR_k. Similarly, we can define the threshold Th^D as

$$Th^D \triangleq \log\frac{Th^d}{\rho\left(1-Th^d\right)} \tag{59}$$

for LSR_k, which is equivalent to the threshold Th^d for p_k. Also, as $Th^d \to 1$, $Th^D \to \infty$.

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, LSR_k, into a ladder variable S_k and a slowly changing term l_k. The definition of a slowly changing variable from Siegmund (2013) is also provided in Appendix A. We define the variable r to be the overshoot of the ladder variable S_{n_d} over a large threshold Th^D at $k = n_d$. Therefore, $r_{n_d} \triangleq S_{n_d} - Th^D$ as $Th^D \to \infty$, and $n_d = \inf\left\{k \geq 1 : S_k \geq Th^D\right\}$. According to the non-linear renewal theory (Siegmund, 2013), the overshoot statistics of LSR_k crossing a large threshold Th^D can be approximated as the statistics of r_{n_d}, provided l_k is slowly changing and $Th^D \to \infty$. 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 1 and the geometric prior distribution of the change point Γ (13), under the two threshold policy Th^s and Th^d, the asymptotic approximate expressions of ADD and FAR as $Th^D \to \infty$ will take the following forms, provided the conditions C1–C4 are satisfied.*

Conditions :

C1: $\{Z_k : k \geq 1\}$ *is nonarithmetic with respect to P_0 and P_1.*

C2: $E_1\left[|Z_1|^2\right]$ *is finite.*

C3: l_k (55) *is a slowly changing variable in the sense defined in Siegmund (2013).*

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

Then,

$$ADD = \frac{Th^D + \bar{r} - \bar{l}}{E_1\left[D\left(f_1^e, f_0\right)\right] + |\log(1-\rho)|} + o(1), \tag{60}$$

and $FAR \approx \frac{\xi}{\rho\exp\left(Th^D\right)}\left(1 + o(1)\right)$, *as* $Th^D \to \infty$, \tag{61}

where

$$\bar{r} = \lim_{n_d \to \infty} E_1\left[r_{n_d}\right], \tag{62}$$

$$\bar{l} = \lim_{k \to \infty} E_1\left[l_k\right], \tag{63}$$

$$\xi = \lim_{n_d \to \infty} E_1\left[\exp\left(-r_{n_d}\right)\right]. \tag{64}$$

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, respectively. $E_1\left[D\left(f_1^e, f_0\right)\right]$ is the expected KLD between the distributions $f_{1,j}^e\left(\cdot|\cdot\right)$ and $f_{2,j}\left(\cdot|\cdot\right)$, and $f_{1,j}^e\left(\cdot|\cdot\right) = f_{1,j}\left(\cdot|\cdot\right)$ 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, $E_0\left[D\left(f_0, f_1^e\right)\right]$ is the expected KLD between f_0 and f_1^e, 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 B. □

Remark 4. As given in Lemma 1 and Theorem 2 of Naha et al. (2021b), $E_1\left[D\left(f_1^e, f_0\right)\right]$ will take the following form,

$$E_1\left[D\left(f_1^e, f_0\right)\right] = \frac{1}{2}\left\{tr\left(\Sigma_0^{-1}\Sigma_{\widetilde{\gamma}}\right) - m - \log\frac{|\mathbf{Q}_a|}{|\Sigma_0|}\right\}, \tag{65}$$

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

$$\Sigma_{\widetilde{\gamma}} = \mathbf{E}_{zz}(0) - \mathbf{C}(\mathbf{A}+\mathbf{BL})\mathbf{E}_{xz}(-1)$$

$$- \left[\mathbf{C}(\mathbf{A}+\mathbf{BL})\mathbf{E}_{xz}(-1)\right]^T + \mathbf{CB}\Sigma_e\mathbf{B}^T\mathbf{C}^T$$

$$+ \mathbf{C}(\mathbf{A}+\mathbf{BL})\Sigma_{x^F z}(\mathbf{A}+\mathbf{BL})^T\mathbf{C}^T$$

$$+ \mathbf{C}(\mathbf{A}+\mathbf{BL})\Sigma_{x^F e}(\mathbf{A}+\mathbf{BL})^T\mathbf{C}^T, \tag{66}$$

$$\text{where } \mathbf{E}_{xz}(-1) = \sum_{i=0}^{\infty}\mathcal{A}^i\mathbf{K}\mathbf{A}_a^{i+1}\mathbf{E}_{zz}(0) \tag{67}$$

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

$$\mathcal{A}\Sigma_{x^F z}\mathcal{A}^T - \Sigma_{x^F z} + \mathbf{K}\mathbf{E}_{zz}(0)\mathbf{K}^T + \mathcal{A}\mathbf{E}_{xz}(-1)\mathbf{K}^T$$

$$+ \left(\mathcal{A}\mathbf{E}_{xz}(-1)\mathbf{K}^T\right)^T = 0, \text{ and} \tag{68}$$

$$\mathcal{A}\Sigma_{x^F e}\mathcal{A}^T - \Sigma_{x^F e} + (\mathbf{I}_n - \mathbf{KC})\mathbf{B}\Sigma_e\mathbf{B}^T(\mathbf{I}_n - \mathbf{KC})^T = 0. \tag{69}$$

Here $\mathcal{A} = (\mathbf{I}_n - \mathbf{KC})(\mathbf{A}+\mathbf{BL})$, which is assumed to be strictly stable. $\mathbf{I}_n$ 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., $\bar{r}$ and ξ, are not dependent on the exact values of the thresholds as long as 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,*

$$ADD \approx \frac{Th^D}{E_1\left[D\left(f_1^e, f_0\right)\right] + |\log(1-\rho)|}, \ as \ Th^D \to \infty \tag{70}$$

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 ξ 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 ξ 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 ξ, 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 $\mathbf{A}_a$ and $\mathbf{Q}_a$. In practice, it is highly likely that the attacker's system parameters $\mathbf{A}_a$ and $\mathbf{Q}_a$ may not be known a priori. In such a case, the attacker's system parameters $\mathbf{A}_a$ and $\mathbf{Q}_a$ can be estimated online from the received observations (true or fake) by fitting a vector autoregressive model to the observations (Akaike, 1969). 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 $\mathbf{A}_a$ and $\mathbf{Q}_a$ will represent the healthy plant model. We have conducted some preliminary studies using a MISO system where our 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 Xie, Zou, Xie, and Veeravalli (2021). Additionally, we also comment that under a replay attack, $\mathbf{A}_a$ and $\mathbf{Q}_a$ can be derived from the normal system model as discussed in Naha et al. (2022b).

4.2. Approximate expression of ANW and ΔLQG

Following similar steps as in Banerjee and Veeravalli (2012), the ANW can be approximated as follows,

$$ANW \approx \frac{E_0\left[t_1\left(Th^S\right)\right]}{E_0\left[t_1\left(Th^S\right)\right] + E_0\left[t_2\left(LSR_{Th^S}, Th^S\right)\right]} \\ \times P\left\{t\left(Th^S\right) < \Gamma\right\}. \tag{71}$$

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. $t_2\left(LSR_{Th^S}, Th^S\right)$ denotes the time interval between the time instances when LSR_k starts from LSR_{Th^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
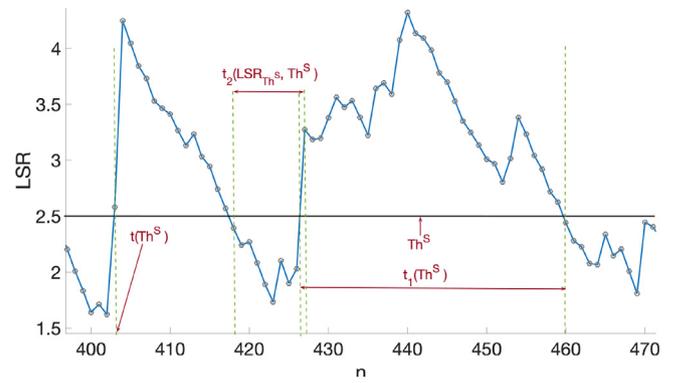
Fig. 6. An example plot of LSR vs. time index n.

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, ΔLQG, is related to ANW as*

$$\Delta LQG = \rho ANW tr\left(H\Sigma_e\right), \tag{72}$$

where $\mathbf{H} = \mathbf{B}^T\Sigma_L\mathbf{B} + \mathbf{U}$ \hfill (73)

and Σ_L *is the solution to the following Lyapunov equation.*

$$(\mathbf{A} + \mathbf{BL})^T \Sigma_L (\mathbf{A} + \mathbf{BL}) - \Sigma_L + \mathbf{L}^T\mathbf{UL} + \mathbf{W} = 0. \tag{74}$$

Proof. The proof of Theorem 2 can be found in Appendix D of Naha et al. (2022a). It is omitted from this paper due to space limitations. □

Theorem 2 shows that ΔLQG is proportional to ANW and a linear function of the watermarking signal variance Σ_e. If watermarking is added at all the time instants, ρANW will become unity, and Theorem 2 will coincide with the special case of always present watermarking as stated in Theorem 3 in Naha et al. (2021b).

4.3. Comparative analysis

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

4.3.1. Comparison with PW-Σ_e

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

The increase in the LQG control cost for PW-Σ_e is as follows, see Theorem 3 from Naha et al. (2021b),

$$\Delta LQG_A = tr\left(H\Sigma_e\right). \tag{75}$$

By comparing the increase in the LQG control cost between the two methods, we can write

$$\widetilde{\Delta LQG} = \Delta LQG_A - \Delta LQG_P = (1 - \rho ANW)\, \Delta LQG_A, \tag{76}$$

where ΔLQG_P denotes the increase in the LQG control cost for the proposed method. Since, $(1 - \rho ANW) < 1$, we can make Claim 1.

Claim 2. *The increase in ADD for the proposed optimal policy with respect to PW-Σ_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, $\bar{l}$, in the ADD expression (60), the rest of the components will be the same for the proposed method and PW-Σ_e. Therefore, the increase in the ADD for the proposed method will be as follows,

$$\Delta ADD = ADD_P - ADD_A \approx \frac{\bar{l}_A - \bar{l}_P}{E_1\left[D\left(f_1^e, f_0\right)\right] + |\log(1-\rho)|}, \tag{77}$$

where the subscripts A and P denote the PW-Σ_e and the proposed method, respectively. $o(1)$ notation is dropped for simplicity. Here $\bar{l}_P$ is the same as given by (55) and (63), and $\bar{l}_A$ will take the following form

$$\bar{l}_A = \lim_{k \to \infty} E_1\left[l_{A,k}\right], \text{ and}$$

$$l_{A,k} = \log\left(SR_0 + \sum_{i=1}^{k}(1-\rho)^{i-1}\mathcal{L}_{b,i}\exp(-Z_i)\right). \tag{78}$$

Since, $\bar{l}_A$ and $\bar{l}_P$ 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-Σ_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_d} will be the same for both methods. Therefore, we can make Claim 3.

4.3.2. Comparison with PW-ΔLQG

Claim 4. *The watermarking signal power for the proposed optimal policy, Σ_{eP}, will be greater than or equal to the watermarking signal power of PW-ΔLQG, Σ_{eB}.*

Since the increase in the LQG control cost is taken to be the same for both the methods, the watermarking signal powers Σ_{eB} and Σ_{eP} for the method PW-ΔLQG and the proposed method, respectively, will be different. The relationships between the watermarking signal power and ΔLQG for both the methods are given as,

$$\Delta LQG = tr\left(H\Sigma_{eB}\right) = \rho ANW tr\left(H\Sigma_{eP}\right). \tag{79}$$

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-ΔLQG.*

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 $E_1\left[D\left(f_1^e, f_0\right)\right]_B$ and $E_1\left[D\left(f_1^e, f_0\right)\right]_P$ as follows.

$$E_1\left[D\left(f_1^e, f_0\right)\right]_P - E_1\left[D\left(f_1^e, f_0\right)\right]_B$$
$$\approx \frac{1}{2}\left(tr\left(\Sigma_\gamma^{-1}\left(\Sigma_{\tilde{\gamma}_P} - \Sigma_{\tilde{\gamma}_A}\right)\right)\right). \tag{80}$$

Here the subscripts B and P denote the method PW-ΔLQG and the proposed method, respectively. By examining (66), we can say $\Sigma_{\tilde{\gamma}_P} - \Sigma_{\tilde{\gamma}_B} \geq 0$. Therefore, from (80) we can write $E_1\left[D\left(f_1^e, f_0\right)\right]_P - E_1\left[D\left(f_1^e, f_0\right)\right]_B \geq 0$, and we can further make Claim 5.

4.4. Optimum Σ_e

Theorem 1 and Corollary 1.1 imply that the increase in KLD will reduce ADD. Therefore, we derive the optimum Σ_e that will maximize KLD for a given fixed upper limit on ΔLQG, denoted as ΔLQG_P for the proposed method. The optimization problem is defined as follows.

$$\max_{\Sigma_e} E_1\left[D\left(f_1^e, f_0\right)\right],$$
$$\text{s.t. } \Delta LQG_P \leq J, \tag{81}$$
$$\Sigma_e \geq 0,$$

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-Σ_e. Moreover, ΔLQG_P (72) is just a scaled version of ΔLQG_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 Σ_e value. Now, the optimization problem for the proposed method becomes identical to the optimization problem for the method PW-Σ_e. According to Theorem 4 from Naha et al. (2021b), the optimum Σ_e for PW-Σ_e will be a rank one positive semi-definite matrix. Therefore, the optimization problem in (81) can be written as

$$\max_{\mathbf{v}_\lambda} E_1\left[D\left(f_1^e, f_0\right)\right]$$
$$\text{s.t. } \Delta LQG_A \leq J_A, \tag{82}$$

where $\mathbf{v}_\lambda = \sqrt{\sigma_e}\mathbf{v}_e$, σ_e is the non-zero eigenvalue of Σ_e and $\mathbf{v}_e$ is the corresponding eigenvector. As discussed in Naha et al. (2021b), the maximization of $E_1\left[D\left(f_1^e, f_0\right)\right]$ with respect to $\mathbf{v}_\lambda$ is same as maximizing the following function,

$$\max_{\mathbf{v}_\lambda} \mathbf{v}_\lambda^T \mathbf{H}_{KLD}\mathbf{v}_\lambda$$
$$\text{s.t. } \Delta LQG_A \leq J_A, \text{ where} \tag{83}$$

$$\mathbf{H}_{KLD} = \mathbf{B}^T\left(\mathbf{I}_n - \mathbf{KC}\right)^T \kappa_e \left(\mathbf{I}_n - \mathbf{KC}\right)\mathbf{B} + \mathbf{B}^T\mathbf{C}^T\mathbf{CB}. \tag{84}$$

Here, κ_e is the solution to the Lyapunov equation

$$\mathcal{A}^T \kappa_e \mathcal{A} - \kappa_e + \left(\mathbf{A} + \mathbf{BL}\right)^T \mathbf{C}^T\mathbf{C}\left(\mathbf{A} + \mathbf{BL}\right) = 0. \tag{85}$$

Since the matrix $\mathcal{A}$ is assumed to be strictly stable, the Lyapunov equation of (85) will have a unique solution. As discussed in Naha et al. (2021b), the optimization problem of (83) can be solved by various methods available in the literature, such as sequential quadratic programming (SQP) (Boggs & Tolle, 1995), interior point method (Forsgren, Gill, & Wright, 2002), simple gradient-based method (Naha et al., 2021b), etc. Interested readers are referred to Naha et al. (2021b) for a detailed analysis, but the same has been removed from the current paper due to space constraints.

5. 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 C provides the required parameters for simulations associated with System A and B.

5.0.1. Optimal policy

Fig. 7 shows the optimal decision variable $\mathbf{u}_{d,k}^*$ vs. p_k plots for three different values of λ_e and a fixed λ_f for System-A. The watermarking signal variance is taken to be a diagonal matrix with equal signal power, σ_e^2. We observe that the optimal policy is a two threshold policy, which validates the theory presented in Section 3.3. A higher λ_e means a stricter constraint on how much
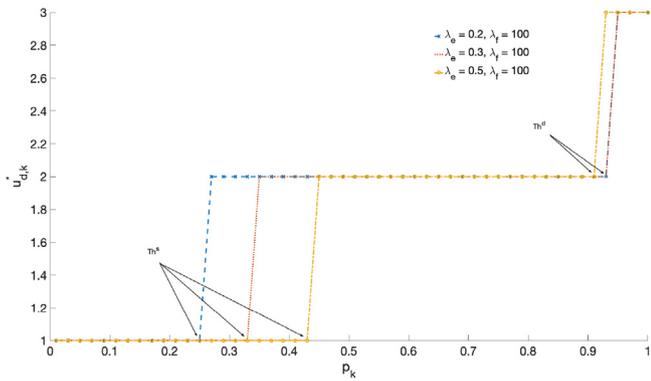
Fig. 7. Optimal policy for different λ_e and fixed λ_f for System-A. $\sigma_e^2 = 1.19$.

Fig. 8. Optimal policy for different λ_f and fixed λ_e for System-A. $\sigma_e^2 = 1.19$.

Fig. 9. s_k, d_k and p_k vs. k for a sample trial run for System-A. $\lambda_e = 0.2$, $\lambda_f = 100$ and $\sigma_e^2 = 1.19$.

Fig. 10. ADD and FAR (%) vs. σ_e^2 plots for two different λ_e and $\lambda_f = 100$, for System-A.
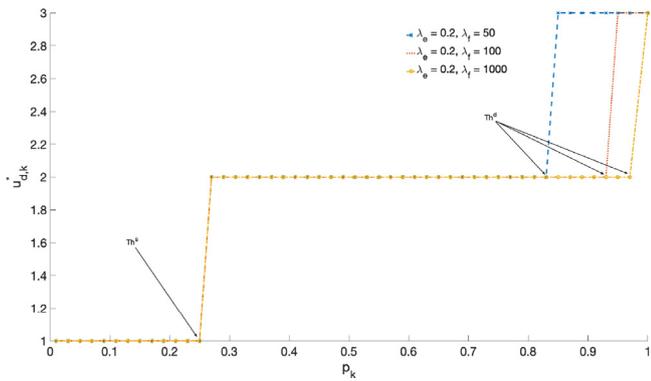
watermarking could be added, which gets reflected into higher Th^s. On the other hand, a higher Th^s means watermarking will be added for fewer samples. As discussed in Section 4.3.1, since the added watermarking has little effect on the FAR, the change in λ_e does not affect the threshold Th^d much.

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

5.0.2. Trial run

Fig. 9 illustrates how the control variables s_k and d_k, and the sufficient statistics p_k change with the time index k for a sample trial run for System-A. The watermarking signal variance is taken to be a diagonal matrix with equal signal power, σ_e^2. 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 $p_k \geq Th^s$, 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, $p_k \geq Th^s$ and watermarking have been added almost all the time after the change point, resulting in faster detection.

5.0.3. ADD and FAR vs. σ_e^2

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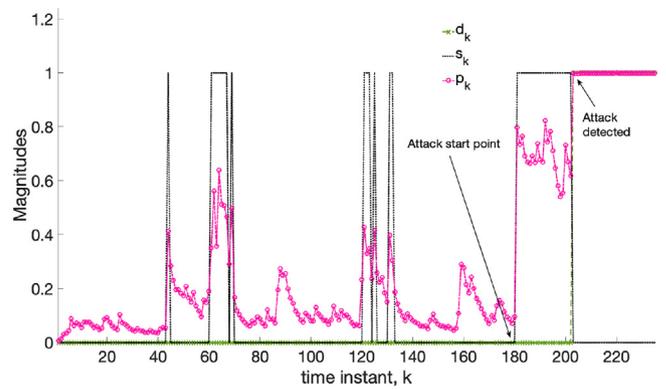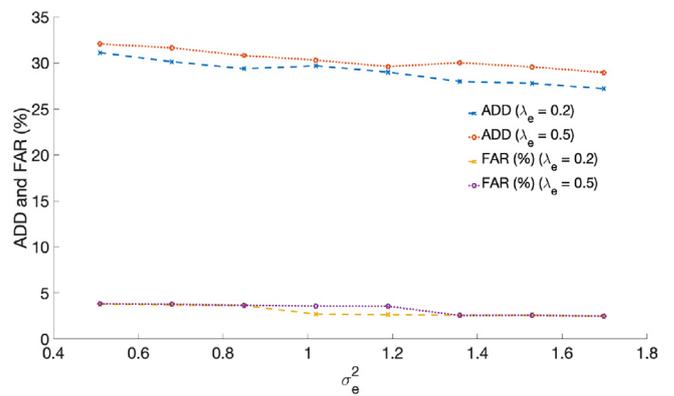
each σ_e^2 point, the thresholds Th^s and Th^d are derived using value iterations from dynamic programming. Then, the ADD and FAR are estimated by MC simulations using the derived thresholds. As discussed before, higher λ_e reduces the usage of watermarking before the attack by increasing the threshold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

Fig. 11 compares the same set of plots as in Fig. 10, but for two different values of λ_f and a fixed λ_e for System-A. As discussed before, the increase in λ_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, whereas FAR reduces. To summarize, ADD increases and FAR decreases with λ_f.

In both the figures, Figs. 10 and 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_{n_d} to some extent, which in turn reduces ξ (64) slightly.
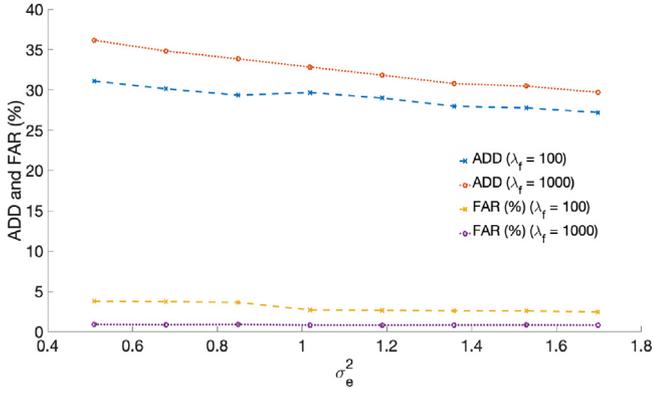
Fig. 11. ADD and FAR (%) vs. σ_e^2 plots for two different λ_f and $\lambda_e = 0.2$, for System-A.
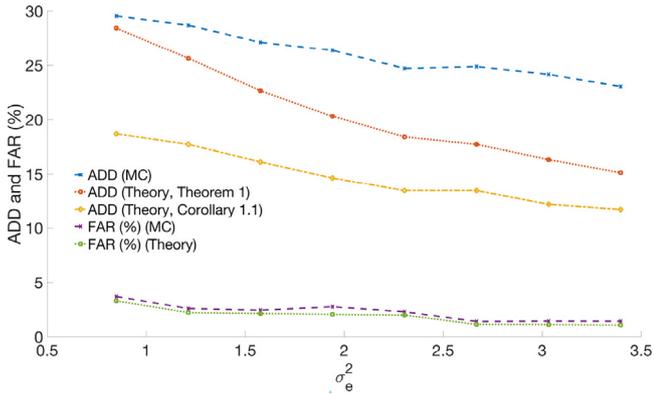
Fig. 13. Comparison between the estimated values and theoretical values. ADD and FAR (%) vs. σ_e^2 plot for System-B. $\lambda_f = 100$ and $\lambda_e = 0.2$.

Fig. 12. Comparison between the estimated values and theoretical values. ADD and FAR (%) vs. σ_e^2 plot for System-A. $\lambda_f = 100$ and $\lambda_e = 0.2$.

Fig. 14. Comparison between the estimated values and theoretical values. ΔLQG vs. σ_e^2 plot for System-A. $\lambda_f = 100$ and $\lambda_e = 0.2$.

5.0.4. Add, FAR and ΔLQG theoretical values

Fig. 12 shows the ADD and FAR vs. σ_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.*, σ_e^2. For each σ_e^2 point, the thresholds Th^s and Th^d are derived using dynamic programming value iterations. Fig. 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 σ_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 σ_e^2 is not making much difference to the estimated ADD. On the other hand, the simulation study shows that ξ (64) does not change much for a small increase in σ_e^2. Therefore, from (61), we can say FAR will only be affected to a small extent due to the increase in σ_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.

Fig. 14 shows the ΔLQG vs. σ_e^2 plot, where ΔLQG 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.*, σ_e^2. From the derived expression of ΔLQG (72), it is evident that the control cost will increase with the increase in watermarking signal power.
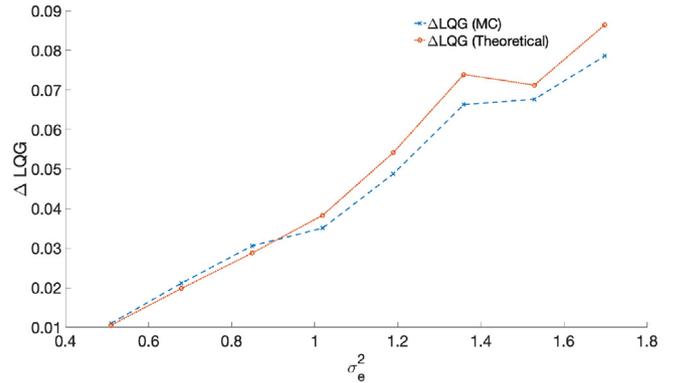
5.0.5. Comparison with PW-Σ_e

Fig. 15 compares the ΔLQG vs σ_e^2 plot from the proposed method and PW-Σ_e assuming a diagonal Σ_e with equal power, σ_e^2, for System-A. For each σ_e^2 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-Σ_e for a fair comparison. From the derived expression of $\widetilde{\Delta LQG}$ (76), we predicted that we would get a large improvement in the control cost since ρ and ANW both are small quantities. Also, the difference will increase with σ_e^2 as ΔLQG_A increases with Σ_e. As predicted from the theory discussed in Section 4.3.1, we observe a large improvement in the control cost (approx. 99% reduction in ΔLQG) for the proposed method in Fig. 15, which validates our Claim 1.

We have shown ADD and FAR vs. σ_e^2 plots for the proposed method and PW-Σ_e using the same parameters as Fig. 15 in Fig. 16. From the derived expression of ΔADD (77), we can comment that the proposed method will take a longer time on average to detect the attack compared to PW-Σ_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 other hand, as discussed in Section 4.3.1, FAR will be the same for both the methods and the same is observed in Fig. 16. To summarize, Fig. 16 supports our Claims 2 and 3.
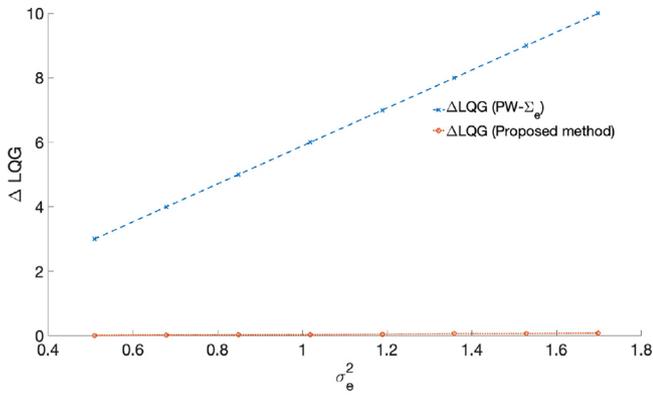
Fig. 15. Comparison between proposed method and PW-Σ_e. ΔLQG vs. σ_e^2 plot for System-A. $\lambda_f = 100$ and $\lambda_e = 0.2$.
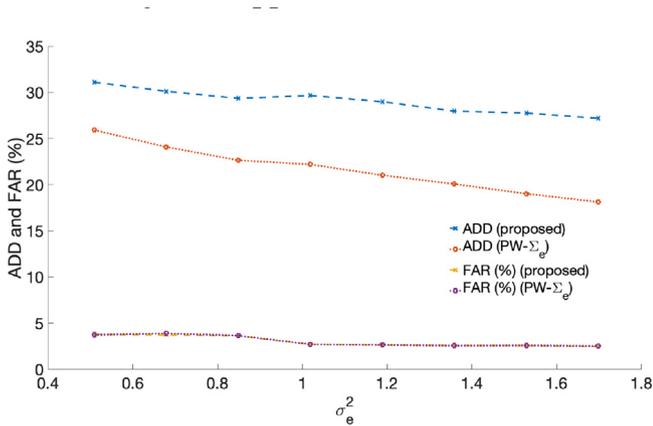
Fig. 16. Comparison between proposed method and PW-Σ_e. ADD and FAR vs. σ_e^2 plot for System-A. $\lambda_f = 100$ and $\lambda_e = 0.2$.

Fig. 17. Comparison between proposed method and PW-ΔLQG. ADD and FAR vs. ΔLQG plot for System-A. $\lambda_f = 100$ and $\lambda_e = 0.3$.

Fig. 18. Comparison between diagonal Σ_e and optimal Σ_e^*. ADD and FAR vs. ΔLQG plot for System-A. $\lambda_f = 100$ and $\lambda_e = 0.3$.

5.0.6. Comparison with PW-ΔLQG

We have shown ADD and FAR vs. ΔLQG plots derived from MC simulations for the proposed method and PW-ΔLQG assuming a diagonal Σ_e with equal power σ_e^2 in Fig. 17 for System-A. For each ΔLQG 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-ΔLQG for a fair comparison. In general, for the proposed method, Th^s decreases and Th^d increases with the increase in ΔLQG or σ_e^2 for fixed λ_e and λ_f. Since the same thresholds are used for PW-ΔLQG, the ADD increases with ΔLQG in the plot. As discussed in Section 4.3.2, 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-ΔLQG. 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 Claims 4 and 5.

5.0.7. Optimum Σ_e

As discussed in Section 4.4, the optimum Σ_e^* reduces the KLD for a fixed upper bound on the ΔLQG, which in turn reduces the ADD. We compare the ADD for the optimum Σ_e^* and the diagonal Σ_e in Fig. 18 for System-A. For each σ_e^2 point, the thresholds Th^s and Th^d are derived using dynamic programming value iterations for the diagonal Σ_e case, and the same thresholds are used for the optimum Σ_e^* case for a fair comparison. We observe an average

increase of 14% (approx.) in the estimated ADD for the optimal Σ_e^*. For the optimal Σ_e^*, the watermarking signal power is mostly concentrated in one eigenvector direction, which results in higher overshoot and a lower FAR.

5.0.8. 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 Fang et al. (2020) for our problem formulation. To fairly compare both methods, we have evaluated ADD and FAR by MC simulations for the same ΔLQG 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 ΔLQG 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 ΔLQG value. Since the periodic watermarking scheme does use any existing evidence extracted from the set Ψ_k of all available information up to the kth 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.

6. 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
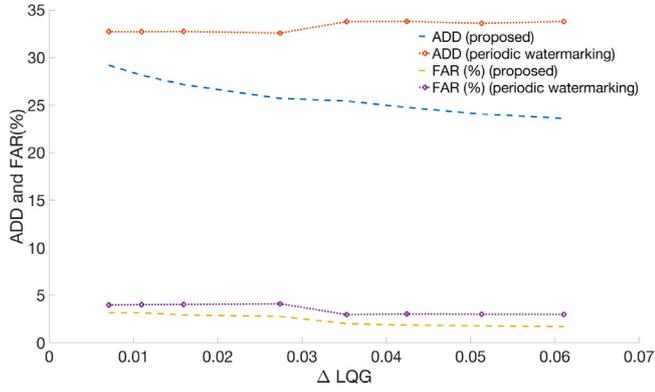
Fig. 19. Comparison between proposed method and a periodic watermarking scheme. ADD and FAR vs. ΔLQG plot for System-A. $\lambda_f = 100$ and $\lambda_e = 0.3$.

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 optional 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 ΔLQG and its relationship with ANW is also derived. Theoretical and simulation studies reveal significant improvement in reducing ΔLQG with a relatively small increase in ADD compared to the method PW-Σ_e, where watermarking is always present. The proposed method is also compared with PW-ΔLQG, where both the methods have the same ΔLQG 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 maximizes the KLD, which will improve the ADD.

Appendix A. Proof of Lemma 3

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.

$$LSR_k = \lambda_k + k|\log(1-\rho)|$$
$$+ \log\left(LSR_0 + \sum_{i=1}^{k}(1-\rho)^{i-1}\mathcal{L}_{d,i}\exp(-\lambda_i)\mathbb{1}_{\{LSR_i < Th^S\}} \right.$$
$$\left. + \sum_{i=1}^{k}(1-\rho)^{i-1}\mathcal{L}_{b,i}\exp(-\lambda_i)\mathbb{1}_{\{LSR_i \geq Th^S\}} \right), \text{ where} \quad (86)$$

$$\lambda_k = \sum_{i=1}^{k}\log\left(\mathcal{L}_{a,i}\right)\mathbb{1}_{\{LSR_i \geq Th^S\}} + \sum_{i=1}^{k}\log\left(\mathcal{L}_{c,i}\right)\mathbb{1}_{\{LSR_i < Th^S\}}. \quad (87)$$

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 λ_k by adding and subtracting $\sum_{i=1}^{k}\log\left(\mathcal{L}_{a,i}\right)\mathbb{1}_{\{LSR_i < Th^S\}}$ to the right hand side of (87) as follows,
$\lambda_k = Z_k + \sum_{i=1}^{k}\log\left(\mathcal{L}_{c,i}\right)\mathbb{1}_{\{LSR_i < Th^S\}} - \sum_{i=1}^{k}\log\left(\mathcal{L}_{a,i}\right)\mathbb{1}_{\{LSR_i < Th^S\}}$,
where Z_k is given in (56). Replacing the first λ_k in (86) by the above expression of λ_k 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 (Siegmund, 2013):

C1: $k^{-1}\max\{|l_1|,\ldots,|l_k|\} \to 0, k \to \infty \quad (88)$

and for every $\epsilon > 0$, there exists k^* and $\delta > 0$, such that for all $k \geq k^*$

C2: $P\left\{ \max_{1 \leq i \leq k\delta} |l_{k+i} - l_k| > \epsilon \right\} < \epsilon. \quad (89)$

l_k from (55) is represented as the summation of three terms, $l_k = l_{1,k} + l_{2,k} - l_{3,k}$, where $l_{1,k} = \log\left(LSR_0 + \sum_{i=1}^{k}(1-\rho)^{i-1}\mathcal{L}_{d,i} \right.$
$\left. \exp(-\lambda_i)\mathbb{1}_{\{LSR_i < Th^S\}}, + \sum_{i=1}^{k}(1-\rho)^{i-1}\mathcal{L}_{b,i}\exp(-\lambda_i)\mathbb{1}_{\{LSR_i \geq Th^S\}} \right)$,
$l_{2,k} = \sum_{i=1}^{k}\log\left(\mathcal{L}_{c,i}\right)\mathbb{1}_{\{LSR_i < Th^S\}}$, and $l_{3,k} = \sum_{i=1}^{k}\log\left(\mathcal{L}_{a,i}\right)$ $\mathbb{1}_{\{LSR_i < Th^S\}}$. Taking absolute values on both sides of l_k equation we can write,

$$|l_k| \leq |l_{1,k}| + |l_{2,k}| + |l_{3,k}|. \quad (90)$$

After the attack start point, LSR_k will gradually increase on average (from condition $C6$ in Theorem 1), and it will first cross Th^S and then Th^D as $k \to \infty$. LSR_k will remain below Th^S for a relatively short period of time compared to the time it takes to cross Th^D, since $Th^D \to \infty$. 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 (90), we can say l_k will also converge to a finite value l, i.e., $l \leq L1 + L2 + L3$ as $k \to \infty$, which means l_k will satisfy condition C1.

We assume that at $k = k_1$, LSR_k crosses Th^S. Therefore, for $k \geq k_1$, we can write $l_{2,k+i} = l_{2,k} = \sum_{i=1}^{k_1}\log\left(\mathcal{L}_{c,i}\right)\mathbb{1}_{\{LSR_i < Th^S\}}$,
$l_{3,k+i} = l_{3,k} = \sum_{i=1}^{k_1}\log\left(\mathcal{L}_{a,i}\right)\mathbb{1}_{\{LSR_i < Th^S\}}$,
$l_{1,k+i} = \log\left(LSR_0 + \sum_{i=1}^{k_1}(1-\rho)^{i-1}\mathcal{L}_{d,i}\exp(-\lambda_i)\times \mathbb{1}_{\{LSR_i < Th^S\}} \right.$
$\left. + \sum_{j=1}^{k+i}(1-\rho)^{j-1}\mathcal{L}_{b,j}\exp(-\lambda_j)\mathbb{1}_{\{LSR_j \geq Th^S\}} \right)$
$l_{1,k} = \log\left(LSR_0 + \sum_{i=1}^{k_1}(1-\rho)^{i-1}\mathcal{L}_{d,i}\exp(-\lambda_i)\mathbb{1}_{\{LSR_i < Th^S\}} \right.$
$\left. + \sum_{i=1}^{k}(1-\rho)^{i-1}\mathcal{L}_{b,i}\exp(-\lambda_i)\mathbb{1}_{\{LSR_i \geq Th^S\}} \right)$. Therefore,

$$l_{k+i} - l_k = l_{1,k+i} - l_{1,k} \text{ for } k \geq k_1. \quad (91)$$

As mentioned before, $\exp(-\lambda_k) \to 0$ as $k \to \infty$, therefore, we can say $P\left\{ l_{1,k+i} - l_{1,k} \right\} \to 0$ for a sufficiently large k, say k^*, and $k^* \geq k_1$. From (91), for $k \geq k^*$, $P\{|l_{k+i} - l_k| > \epsilon\} = 0$, which in turn will satisfy condition C2.

Appendix B. 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 (56) 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

$$Z_1 = -\frac{1}{2}\log\frac{|\Sigma_{1,1}|}{|\Sigma_0|} - \frac{1}{2}\left(\widetilde{\gamma}_1 - \mu_{1,1}\right)^T \Sigma_{1,1}^{-1}\left(\widetilde{\gamma}_1 - \mu_{1,1}\right)$$
$$+ \frac{1}{2}\widetilde{\gamma}_1^T \Sigma_0^{-1}\widetilde{\gamma}_1 \quad (92)$$

From (38), (40), and (42), we can say that all the elements of (92) are either finite or having Gaussian distributions with finite

means and variances, which in turn ensures that $E_1\left[\mid Z_1\mid^2\right]$ is finite.

Condition C3 is proven in Appendix A. From the expressions of Σ_0 (42) and $\Sigma_{\widetilde{\gamma}}$ (67), we can say that under the practical assumptions of the plant model and attacker's system parameters $0 < E_1\left[D\left(f_1^e, f_0\right)\right] < \infty$ from (65). In a similar way we can also show that $0 < E_1\left[D\left(f_0, f_1^e\right)\right] < \infty$. 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 $k = \Gamma$ the test statistics LSR_k will cross the threshold Th^D at $k = \tau$ for the first time which is equivalent to the test statistics p_k crossing the threshold Th^d. To derive the expression of ADD, we assume, $T_D = \tau - \Gamma$. After adding and subtracting Th^D to (53) and rearranging the terms, it will take the following form at $k = T_D$,

$$S_{T_D} = Th^D + \left(LSR_{T_D} - Th^D\right) - l_{T_D}. \tag{93}$$

According to the nonlinear renewal theory (Siegmund, 2013), the overshoot statistics of $LSR_{T_D} - Th^D$ can be approximated by the overshoot statistics of S_{T_D}, i.e., $r_{n_d} = S_{T_D} - Th^D$, provided $Th^D \to \infty$. Moreover, the slowly changing term $l_k \to l$ as $k \to \infty$, where l is a RV (Tartakovsky & Veeravalli, 2005). Taking expectations on both sides of (93), we get

$$E_1\left[S_{T_D}\right] = Th^D + \bar{r} - \bar{l} + o(1), \tag{94}$$

where $\bar{r} = \lim_{n_d \to \infty} E_1\left[r_{n_d}\right]$ and $\bar{l} = \lim_{k \to \infty} E_1\left[l_k\right]$. The following expression of $E_1\left[S_{T_D}\right]$ is derived by taking expectations on both sides of (54) (Tartakovsky & Veeravalli, 2005),

$$E_1\left[S_{T_D}\right] = E_1\left[T_D\right]\left(E_1\left[Z_1\right] + \mid\log(1 - \rho)\mid\right). \tag{95}$$

Furthermore, $E_1\left[Z_1\right]$ can be approximated as $E_1\left[D\left(f_1^e, f_0\right)\right]$ (65) as explained in Naha et al. (2021b). Combining (94) and (95), and rearranging the term we get (60), where $ADD = E_1\left[T_D\right]$.

A brief derivation of FAR is provided as follows. A detailed one can be found in Tartakovsky and Veeravalli (2005). Using (43), we can write

$$FAR = E^\pi\left[1 - p_\tau\right] = E^\pi\left[\frac{1}{\exp\left(LSR_\tau\right)}\frac{1}{\rho + \exp\left(-LSR_\tau\right)}\right] \tag{96}$$

False alarm will occur when LSR_k crosses Th^D during the normal system operation. Therefore, $\exp\left(-LSR_\tau\right) \leq \exp\left(-Th^D\right) \to 0$ as $Th^D \to \infty$. So, (96) can be approximated for $Th^D \to \infty$ as

$$FAR = \frac{1}{\rho}\exp\left(-Th^D\right) E^\pi\left[\exp\left(Th^D - LSR_\tau\right)\right](1 + o(1)) \tag{97}$$

$E^\pi\left[\exp\left(Th^D - LSR_\tau\right)\right]$ can be approximated by ξ using the overshoot r_{n_d} statistics (Siegmund, 2013) as given in (64). Replacing $E^\pi\left[\exp\left(Th^D - LSR_\tau\right)\right]$ by ξ in (97), we get (61).

Appendix C. System parameters

The following system parameters are used for simulation study. $\rho = 0.001$.

System-A parameters:

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

$$\mathbf{Q} = diag\begin{bmatrix} 1 & 1 \end{bmatrix}, \ \mathbf{R} = 1, \ \mathbf{W} = diag\begin{bmatrix} 1 & 2 \end{bmatrix},$$

$$\mathbf{U} = diag\begin{bmatrix} 0.4 & 0.7 \end{bmatrix}, \ \mathbf{A}_a = 0.5, \ and \ \mathbf{Q}_a = 7.5.$$

System-B parameters:

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

$$\mathbf{C} = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \end{bmatrix}, \ \mathbf{R} = diag\begin{bmatrix} 0.5 & 0.5 \end{bmatrix},$$

$$\mathbf{Q} = diag\begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix}, \ \mathbf{U} = diag\begin{bmatrix} 2 & 2 \end{bmatrix},$$

$$\mathbf{W} = diag\begin{bmatrix} 5 & 5 & 1 & 1 \end{bmatrix}, \ \mathbf{Q}_a = diag\begin{bmatrix} 6 & 6 \end{bmatrix},$$

$$\mathbf{A}_a = diag\begin{bmatrix} 0.4 & 0.1 & 0.1 & 0.7 \end{bmatrix}.$$

References

Abrams, M., & Weiss, J. (2008). *Malicious control system cyber security attack case study – Maroochy Water Services, Australia, Vol. 253* (pp. 73–82). USA: MITRE Corp, no. August.

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics, 21*(1), 243–247.

Alguliyev, R., Imamverdiyev, Y., & Sukhostat, L. (2018). Cyber-physical systems and their security issues. *Computers in Industry, 100*(July 2017), 212–223, [Online]. Available: https://doi.org/10.1016/j.compind.2018.04.017.

Arnold, F., Hermanns, H., Pulungan, R., & Stoelinga, M. (2014). Time-dependent analysis of attacks. In *International conference on principles of security and trust* (pp. 285–305). Springer.

Banerjee, T., & Veeravalli, V. V. (2012). Data-efficient quickest change detection with on-off observation control. *Sequential Analysis, 31*(1), 40–77.

Bertsekas, D. P. (1995). *Dynamic programming and optimal control, Vol. 1.* Belmont, MA: Athena scientific, no. 2.

Beutler, F. J., & Ross, K. W. (1985). Optimal policies for controlled Markov chains with a constraint. *Journal of Mathematical Analysis and Applications, 112*(1), 236–252.

Boggs, P. T., & Tolle, J. W. (1995). Sequential quadratic programming. *Acta Numerica, 4*(1995), 1–51.

Cardenas, A., Amin, S., Sinopoli, B., Giani, A., Perrig, A., Sastry, S., et al. (2009). Challenges for securing cyber physical systems. In *Workshop on future directions in cyber-physical systems security, Vol. 5.* no. 1.

Chen, Y., Kar, S., & Moura, J. M. (2018). Cyber-Physical attacks with control objectives. *IEEE Transactions on Automatic Control, 63*(5), 1418–1425.

Du, D., Li, X., Li, W., Chen, R., Fei, M., & Wu, L. (2019). ADMM-based distributed state estimation of smart grid under data deception and denial of service attacks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 49*(8), 1698–1711.

Fang, C., Qi, Y., Cheng, P., & Zheng, W. X. (2020). Optimal periodic watermarking schedule for replay attack detection in cyber–physical systems. *Automatica, 112*, Article 108698.

Fawzi, H., Tabuada, P., & Diggavi, S. (2014). Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE Transactions on Automatic Control, 59*(6), 1454–1467.

Ferrari, R. M., & Teixeira, A. M. (2017). Detection and isolation of replay attacks through sensor watermarking. *IFAC-PapersOnLine, 50*(1), 7363–7368.

Forsgren, A., Gill, P. E., & Wright, M. H. (2002). Interior methods for nonlinear optimization. *SIAM Review, 44*(4), 525–597.

Forti, N., Battistelli, G., Chisci, L., Li, S., Wang, B., & Sinopoli, B. (2018). Distributed joint attack detection and secure state estimation. *IEEE Transactions on Signal and Information Processing over Networks, 4*(1), 96–110.

Fuh, C. D., & Tartakovsky, A. G. (2019). Asymptotic Bayesian theory of quickest change detection for hidden markov models. *IEEE Transactions on Information Theory, 65*(1), 511–529.

Ge, X., Han, Q. L., Zhong, M., & Zhang, X. M. (2019). Distributed Krein space-based attack detection over sensor networks under deception attacks. *Automatica, 109*, Article 108557.

Jonsson, E., & Olovsson, T. (1997). A quantitative model of the security intrusion process based on attacker behavior. *IEEE Transactions on Software Engineering, 23*(4), 235–245.

Ko, W. H., Satchidanandan, B., & Kumar, P. R. (2019). Dynamic watermarking-based defense of transportation cyber-physical systems. *ACM Transactions on Cyber-Physical Systems, 4*(1).

Langner, R. (2011). Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security & Privacy, 9*(3), 49–51.

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

Liu, Y.-C., Bianchin, G., & Pasqualetti, F. (2020). Secure trajectory planning against undetectable spoofing attacks. *Automatica, 112*, Article 108655.

Lorden, G., et al. (1971). Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics, 42*(6), 1897–1908.

Mo, Y., Chabukswar, R., & Sinopoli, B. (2014). Detecting integrity attacks on SCADA systems. *IEEE Transactions on Control Systems Technology*, 22(4), 1396–1407.

Mo, Y., & Sinopoli, B. (2009). Secure control against replay attacks. In *2009 47th annu. allert. conf. commun. control. comput. allert. 2009* (pp. 911–918). IEEE.

Mo, Y., Weerakkody, S., & Sinopoli, B. (2015). Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs. *IEEE Control Systems*, 35(1), 93–109.

Mousavinejad, E., Yang, F., Han, Q. L., & Vlacic, L. (2018). A novel cyber attack detection method in networked control systems. *IEEE Transactions on Cybernetics*, 48(11), 3254–3264.

Naha, A., Teixeira, A., Ahlen, A., & Dey, S. (2021a). Deception attack detection using reduced watermarking. In *2021 European control conference (ECC)* (pp. 74–80). IEEE.

Naha, A., Teixeira, A., Ahlén, A., & Dey, S. (2021b). Quickest detection of deception attacks in networked control systems with physical watermarking. arXiv preprint arXiv:2101.01466.

Naha, A., Teixeira, A., Ahlén, A., & Dey, S. (2022a). Quickest detection of deception attacks on cyber-physical systems with a parsimonious watermarking policy. arXiv preprint arXiv:2201.09389.

Naha, A., Teixeira, A. M. H., Ahlen, A., & Dey, S. (2022b). Sequential detection of Replay attacks. *IEEE Transactions on Automatic Control*, 1.

Naha, A., Teixeira, A., Ahlén, A., & Dey, S. (2022c). Sequential detection of replay attacks with a parsimonious watermarking policy. In *American control conference*. Accepted.

Park, G., Lee, C., Shim, H., Eun, Y., & Johansson, K. H. (2019). Stealthy adversaries against uncertain Cyber-Physical Systems: Threat of robust zero-dynamics attack. *IEEE Transactions on Automatic Control*, 64(12), 4907–4919.

Pasqualetti, F., Dorfler, F., & Bullo, F. (2013). Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, 58(11), 2715–2729.

Premkumar, K., & Kumar, A. (2008). Optimal sleep-wake scheduling for quickest intrusion detection using sensor networks. In *Proc. - IEEE INFOCOM, no. ii* (pp. 2074–2082).

Prochaska, B. (1973). A note on the relationship between the geometric and exponential distributions. *The American Statistician*, 27(1), 27.

Salimi, S., Dey, S., & Ahlen, A. (2019). Sequential detection of deception attacks in networked control systems with watermarking. In *2019 18th European control conference, ECC 2019* (pp. 883–890).

Sánchez, H. S., Rotondo, D., Escobet, T., Puig, V., Saludes, J., & Quevedo, J. (2019). Detection of replay attacks in cyber-physical systems using a frequency-based signature. *Journal of the Franklin Institute*, 356(5), 2798–2824.

Satchidanandan, B., & Kumar, P. R. (2017). Dynamic watermarking: Active defense of networked cyber–physical systems. *Proceedings of the IEEE*, 105(2), 219–240.

Satchidanandan, B., & Kumar, P. R. (2020). On the design of security-guaranteeing dynamic watermarks. *IEEE Control Systems Letters*, 4(2), 307–312.

Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of Probability and its Applications*, 8(1), 22–46.

Siegmund, D. (2013). *Sequential analysis: Tests and confidence intervals*. Springer Science & Business Media.

Tartakovsky, A. G. (2017). On asymptotic optimality in sequential change-point detection: Non-iid case. *IEEE Transaction on Information Theory*, 63(6), 3433–3450.

Tartakovsky, A., Nikiforov, I., & Basseville, M. (2014). *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press.

Tartakovsky, A. G., & Veeravalli, V. V. (2005). General asymptotic Bayesian theory of quickest change detection. *Theory of Probability and its Applications*, 49(3), 458–497.

Trapiello, C., Rotondo, D., Sanchez, H., & Puig, V. (2019). Detection of replay attacks in CPSs using observer-based signature compensation. In *2019 6th int. conf. control. decis. inf. technol. CoDIT 2019* (pp. 1–6). IEEE.

Xie, L., Zou, S., Xie, Y., & Veeravalli, V. V. (2021). Sequential (quickest) change detection: Classical results and new directions. *IEEE Journal on Selected Areas in Information Theory*, 2(2), 494–514.

Yakir, B. (1994). Optimal detection of a change in distribution when the observations form a Markov chain with a finite state space. *Lecture Notes-Monograph Series*, 23, 346–358.

Ye, D., Zhang, T. Y., & Guo, G. (2019). Stochastic coding detection scheme in cyber-physical systems against replay attack. *Information Sciences (Ny)*, 481(61773097), 432–444, [Online]. Available: https://doi.org/10.1016/j.ins.2018.12.091.

Yılmaz, M. H., & Arslan, H. (2015). A survey: Spoofing attacks in physical layer security. In *2015 IEEE 40th local computer networks conference workshops (LCN workshops)* (pp. 812–817). IEEE.

Arunava Naha received his Ph.D. from the Electrical Engineering Department, Indian Institute of Technology Kharagpur, India in 2018. He is currently a visiting researcher at the Electrical Engineering Department, Uppsala University. He has also worked for two and a half years at Samsung R&D Institute India - Bangalore (SRIB) as Chief Engineer. His current research interests include statistical signal processing, signal-based condition monitoring, the security of cyber–physical systems, etc. He is an Associate Editor for IEEE Transactions on Instrumentation and Measurement.

André M.H. Teixeira received the M.Sc. degree in electrical and computer engineering from the Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, in 2009, and the Ph.D. degree in automatic control from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2014.

He is currently an Associate Professor at the Department of Information Technology, Uppsala University, Sweden. Prior to this, he was an Associate Senior Lecturer at the Department of Electrical Engineering, Uppsala University during 2017–2021. From 2015 to 2017, he was an Assistant Professor at the Faculty of Technology, Policy and Management, Delft University of Technology.

Dr. Teixeira serves as Associate Editor for Automatica and he was a recipient for the Best Student-Paper Award from the IEEE Multi-Conference on Systems and Control in 2014, and an Honorable Mention for the Paul M. Frank Award at the IFAC SAFEPROCESS in 2018. He was awarded a Starting Grant by the Swedish Research Council in 2019, and he is among the 20 young researchers in Sweden that received the Future Research Leaders 7 grant by the Swedish Foundation for Strategic Research in 2020. His current research interests include secure and resilient control systems, distributed fault detection, distributed optimization, and large-scale dynamical systems.

Anders Ahlén was born in Kalmar, Sweden. He is currently a Senior Professor of Signal Processing at the Signals and Systems Division, Department of Electrical Engineering, Uppsala University. He received the Ph.D. degree in Automatic Control from Uppsala University. From 1996–2021 he held the Chair in Signal Processing and was the Head of the Signals and Systems Division, Department of Electrical Engineering, Uppsala University. During 1991, he was a Visiting Researcher with the Department of Electrical and Computer Engineering, The University of Newcastle, Australia, where he has been a Visiting Professor several times since 2008. He was also a visiting professor with the University of South Australia, Adelaide, during 2018. From 2001 to 2004, he was the CEO of Dirac Research AB, a company offering state-of-the-art audio signal processing solutions. From 2005 to 2020, he was the Chairman of the Board-of-Directors, and since 2020 he is member of the Board-of-Directors at the same company. His research interests are currently focused on signal processing and machine learning, wireless sensor networks, wireless control, security and privacy of cyber–physical systems, and audio signal processing. From 1998 to 2004, he was the Editor of Signal and Modulation Design for the IEEE TRANSACTIONS ON COMMUNICATIONS.

Subhrakanti Dey received the Ph.D. degree from the Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, in 1996.

He is currently a Professor and Head of the Signals and Systems division in the Dept of Electrical Engineering at Uppsala University, Sweden. He has also held professorial positions at NUI Maynooth, Ireland and University of Melbourne, Australia. His current research interests include wireless communications and networks, signal processing for sensor networks, networked control systems, and molecular communication systems. He is a Senior Editor for IEEE Transactions of Control of Network Systems and IEEE Control Systems Letters, and an Associate Editor for Automatica.