Detecting Integrity Attacks on Control Systems using a Moving Target Approach

Sean Weerakkody   Bruno Sinopoli

Abstract— Maintaining the security of control systems in the presence of integrity attacks is a significant challenge. In literature, several possible attacks against control systems have been formulated including replay, false data injection, and zero dynamics attacks. The detection and prevention of these attacks require the defender to possess a particular subset of trusted communication channels. Alternatively, these attacks can be prevented by keeping the system model secret from the adversary. In this paper, we consider an adversary who has the ability to modify and read all sensor and actuator channels. To thwart this adversary, we introduce external states dependent on the state of the control system, with linear time-varying dynamics unknown to the adversary. We also include sensors to measure these states. The presence of unknown time-varying dynamics is leveraged to detect an adversary who simultaneously aims to identify the system and inject stealthy outputs. Potential attack strategies and bounds on the attacker's performance are provided.

I. INTRODUCTION

Cyber-Physical Systems (CPSs), referring to the tight interconnection of sensing, communication, and control in physical spaces, are becoming widespread in today's society. Indeed, these systems will serve a significant role in several applications including transportation, water distribution, medical technologies, manufacturing, and of course the smart grid. Due to the proliferation of CPSs in critical infrastructures, their safety and security are of paramount importance. There have already been several powerful attacks against CPSs. One major example is Stuxnet, which targeted Supervisory Control and Data Acquisition Systems (SCADA) at uranium enrichment facilities in Iran [1], [2]. Here, the adversary was able to appropriate controllers running centrifuges at the plant, and avoid detection by replaying previous measurements to the system operator. An additional example is the Maroochy Shire incident where a disgruntled employee performed an attack on a SCADA based sewage control system [3].

Previous work [4] has suggested that existing tools in cyber security are insufficient to address attacks on CPSs due to the underlying physical system. Two main classes of attacks defined by [4] are denial of service attacks where an attacker restricts the flow of information between the plant and control center and integrity attacks where an adversary can alter control inputs and sensor outputs. An intelligent adversary can potentially cause physical damage to a system while manipulating sensor measurements to avoid detection. As such, integrity attacks are the main focus of this paper.

Several integrity attacks have been investigated in the literature. For instance, [5], [6] analyze zero dynamic attacks where an adversary injects inputs into both the actuators and sensors so as to bias the state without inserting a net bias on the sensor measurements. False data injection attacks on measurements, where an adversary alters a subset of sensor measurements to induce destabilizing control inputs from the defender have also been studied. Liu et. al. [7] first studied false data injection attacks in the context of electricity grids. Furthermore, in [8], the authors consider false data injection in control systems, providing sufficient and necessary conditions for an attacker to destabilize a system while introducing a bounded error on the outputs. Finally, replay attacks where an adversary repeats a sequence of past measurements are analyzed in [9], [10].

The detection and prevention of integrity attacks on control systems against adversaries who are aware of the system model rely on the presence of one or more secure communication channels between the system operator and the plant. For instance, [6] provides sufficient and necessary conditions for zero dynamic attacks based on the actuators and sensors in possession of the adversary. If the adversary has access to all sensors and actuators a trivial zero dynamics attack is to subtract ones influence from the true measurements. To prevent false data injection attacks in control systems, a particular subset of measurements must be secure from the adversary [8]. Moreover, [11] proposes assigning security indices to each sensor to quantify the effort required for an adversary to introduce a successful false data injection attack. Physical watermarking, used to detect replay attacks in [9], [10] and robust attacks defined in [12] rely on the ability to inject secret noisy inputs into the control system. [13] which considers the problem of robust estimation and control in the presence of integrity attacks, relies on the assumption that the attacker is only able to manipulate less than half the sensors.

In this paper, we consider the scenario where an adversary has access to all communication channels. Thus, to prevent an attack, an adversary must not be aware of the full system model. [14] considers the problem of altering system matrices to avoid zero dynamic attacks. However, in practice an adversary can use his access to both inputs and outputs

to identify the system. Moreover, a malicious insider such as the attacker in the Maroochy Shire incident is likely to be aware of the system model. Consequently, we propose introducing extraneous states correlated to the ordinary states of the system so that modification of the original states will impact the extraneous states. The extraneous states will have linear time-varying dynamics known to the system operator and hidden from the adversary. The dynamics act as a moving target, changing fast enough so the adversary does not have adequate opportunity to identify the extraneous system. In this scenario, we propose attacks for the adversary and obtain detection bounds.

The rest of the paper is organized as follows. In Section II, we introduce our system model and control strategy. In Section III, we propose the moving target approach to detect integrity attacks on control systems. In Section IV, we summarize the attacker's capabilities and propose two attack models. In Section V, we analyze bounds on the attacker's performance. Section VI concludes the paper.

II. System Model

In this section, we introduce the model for our system. In particular, we assume our cyber-physical system can be modeled as a discrete time control system where

$$x_{k+1} = Ax_k + Bu_k + w_k, \tag{1}$$
$$y_k = Cx_k + v_k. \tag{2}$$

Here $x_k \in \mathbb{R}^n$ is the state vector at time k and $u_k \in \mathbb{R}^p$ is a collection of control inputs. A suite of sensors are used to monitor the state. Here $y_k \in \mathbb{R}^m$ is a vector of sensor measurements taken at time k. w_k is the independent and identically distributed (IID) process noise with probability distribution given by $\mathcal{N}(0, Q)$ where $Q \succ 0$. Meanwhile, v_k is the IID measurement noise with distribution given by $v_k \sim \mathcal{N}(0, R)$ where $R \succ 0$. We assume that (A, C) is detectable. Additionally, (A, B) and $(A, Q^{\frac{1}{2}})$ are assumed to be stabilizable.

The set of measurements y_k are sent to the SCADA center in order to compute the optimal control input. For our purposes, we assume that the operator wishes to minimize a quadratic function of the states and inputs as follows

$$J = \lim_{T \to \infty} \frac{1}{T+1} \mathbb{E}\left[\sum_{k=0}^{T} x_k^T W x_k + u_k^T U u_k \right], \tag{3}$$

where $W \in \mathbb{R}^{n \times n}$, $U \in \mathbb{R}^{p \times p}$ are positive definite matrices defining the relative cost of each state and input. The optimal control input for the given cost function is a combination of a Kalman filter and a linear state feedback controller [15].

The Kalman filter computes the minimum mean squared error state estimate $\hat{x}_{k|k}^r$ [1] given the previous set of measurements up to y_k denoted by $y_{1:k}$. We assume that the system has been running for a long time so that the Kalman filter

---

[1] The superscript r is used to distinguish the ordinary state estimate from the state estimate obtained through the moving target model.

has converged to a fixed gain linear estimator.

$$\hat{x}_{k+1|k}^r = A\hat{x}_{k|k}^r + Bu_k, \tag{4}$$
$$\hat{x}_{k|k}^r = (I - KC)\hat{x}_{k|k-1}^r + Ky_k, \tag{5}$$
$$K = PC^T(CPC^T + R)^{-1}, \tag{6}$$
$$P = APA^T + Q - APC^T(CPC^T + R)^{-1}CPA^T. \tag{7}$$

The optimal control input with respect to (3) is given by

$$u_k^* = L\hat{x}_{k|k}^r, \quad L = -(B^T SB + U)^{-1}B^T SA, \tag{8}$$

and S satisfies the following Riccati equation

$$S = A^T SA + W - A^T SB(B^T SB + U)^{-1}B^T SA. \tag{9}$$

A bad data detector can be utilized to determine whether a malicious attack is occurring. Typically the bad data detector can be written as a threshold-based detector where

$$g_k(\mathcal{I}_k) \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \eta_k. \tag{10}$$

Here, $\mathcal{I}_k$ is the information available to the defender. The null hypothesis $\mathcal{H}_0$ is that the system is operating normally while the alternate hypothesis $\mathcal{H}_1$ is that the system is under attack. A more specific detector will be discussed later in the article. We furthermore define the probability of detection β_k and false alarm α as

$$\beta_k = \Pr\left(g_k\left(\mathcal{I}_k\right) > \eta_k | \mathcal{H}_1\right), \ \alpha = \Pr\left(g_k\left(\mathcal{I}_k\right) < \eta_k | \mathcal{H}_1\right). \tag{11}$$

Observe that α is independent of k since the system is stationary. Regardless of the information available to a system operator, an attacker with knowledge of the input to output model, as well as the ability to manipulate sensor measurements and control inputs can generate undetectable attacks. [16].

For instance, an adversary can simply subtract the influence he inserts through the control inputs from the system outputs as follows

$$x_{k+1} = Ax_k + B(u_k^* + u_k^a) + w_k, \tag{12}$$
$$y_k = Cx_k + v_k - s_k^a, \tag{13}$$

where s_k^a is given by

$$x_{k+1}^a = Ax_k^a + Bu_k^a, \tag{14}$$
$$s_k^a = Cx_k^a. \tag{15}$$

In this case, the attack has zero net effect on the outputs and as a result $\beta_k = \alpha$.

III. The Moving Target

As discussed in the previous section, an adversary who is both aware of the system model and has access to all channels can generate undetectable attacks. In this work, we propose introducing linear time-varying dynamics, unknown to the adversary, but known to the defender, into the system. The defender can leverage his knowledge of the system to detect integrity attacks by the adversary. Moreover, by introducing time-varying dynamics, the defender limits the adversary's ability to identify the system using his access to measurements and inputs. The time-varying dynamics act as a moving target.

A. Extended Model

We extend the state x_k to include extraneous states $\tilde{x}_k \in \mathbb{R}^{\tilde{n}}$ as follows

$$\begin{bmatrix} \tilde{x}_{k+1} \\ x_{k+1} \end{bmatrix} = \mathcal{A}_k \begin{bmatrix} \tilde{x}_k \\ x_k \end{bmatrix} + \mathcal{B}_k u_k + \begin{bmatrix} \tilde{w}_k \\ w_k \end{bmatrix}, \quad (16)$$

where

$$\mathcal{A}_k \triangleq \begin{bmatrix} A_{1,k} & A_{2,k} \\ 0 & A \end{bmatrix}, \quad \mathcal{B}_k \triangleq \begin{bmatrix} B_k \\ B \end{bmatrix}. \quad (17)$$

Moreover, we introduce additional sensors $\tilde{y}_k \in \mathbb{R}^{\tilde{m}}$ to measure the extraneous states.

$$\begin{bmatrix} \tilde{y}_k \\ y_k \end{bmatrix} = \mathcal{C}_k \begin{bmatrix} \tilde{x}_k \\ x_k \end{bmatrix} + \begin{bmatrix} \tilde{v}_k \\ v_k \end{bmatrix}, \quad \mathcal{C}_k \triangleq \begin{bmatrix} C_k & 0 \\ 0 & C \end{bmatrix}. \quad (18)$$

The matrices are assumed to be IID random variables which are independent of the sensor and process noise with distribution

$$A_{1,k}, A_{2,k}, B_k, C_{k+1} \sim f_{A_{1,k}, A_{2,k}, B_k, C_{k+1}}(A_1, A_2, B, C). \quad (19)$$

Furthermore, we also assume that

$$\begin{bmatrix} \tilde{w}_k \\ w_k \end{bmatrix} \sim \mathcal{N}(0, \mathcal{Q}), \quad \begin{bmatrix} \tilde{v}_k \\ v_k \end{bmatrix} \sim \mathcal{N}(0, \mathcal{R}), \quad (20)$$

where

$$\mathcal{Q} = \begin{bmatrix} \tilde{Q} & \tilde{Q}_{12} \\ \tilde{Q}_{12}^T & Q \end{bmatrix} \succ 0, \quad \mathcal{R} = \begin{bmatrix} \tilde{R} & \tilde{R}_{12} \\ \tilde{R}_{12}^T & R \end{bmatrix} \succ 0. \quad (21)$$

Remark 1: While we assume the structure of the system introduced above with IID matrices $A_{1,k}, A_{2,k}, B_k, C_{k+1}$, the moving target design can still be effective in other scenarios. For instance, the dynamics need not be linear as long as the defender can accurately model the system. Moreover, the system parameters do not have to evolve at each time step though the longer the target remains in place, the easier it is for the adversary to identify the system. In addition, the matrices $A_{1,k}$, $A_{2,k}$, or B_k may be zero matrices. However, at least one of $A_{2,k}$, or B_k must be nonzero to ensure $\tilde{x}_k$ is not decoupled from x_k.

Remark 2: The defender must be able to introduce extraneous states with time-varying dynamics correlated to the original state of the system. The extraneous states are application dependent and are to be decided by the system operator. Nonetheless, the system operator can leverage existing waste products of the system, for instance the heat dissipated by a reaction or process. The dynamics can be made time-varying by changing conditions at the plant. Alternatively, the defender can introduce dynamics into the system. For instance, the defender can introduce RLC circuits which measure the states. Time varying dynamics can be incorporated by including variable resistors or capacitors. By varying the components of the circuit according to some IID distribution at each time step, the defender can generate IID system matrices.

Remark 3: In the above formulation we assume that the defender is aware of the real time system matrices although they are random. In general, this information should not be sent over the network since doing so amounts to the existence of a secure communication channel. The secure communication channel could be leveraged to detect an attack without considering a moving target approach, for instance through physical watermarking [12]. Alternatively, we can generate pseudo random system matrices using a seed for a random number generator. In this case, the seed will act as the secret for the system.

B. Estimation and Detection

The presence of additional sensors allows us to improve our estimate of the state. In particular, we can incorporate an additional Kalman filter to estimate the state as follows.

$$\begin{bmatrix} \hat{\tilde{x}}_{k+1|k} \\ \hat{x}_{k+1|k} \end{bmatrix} = \mathcal{A}_k \left((I - \mathcal{K}_k \mathcal{C}_k) \begin{bmatrix} \hat{\tilde{x}}_{k|k-1} \\ \hat{x}_{k|k-1} \end{bmatrix} + \mathcal{K}_k \begin{bmatrix} \tilde{y}_k \\ y_k \end{bmatrix} \right) + \mathcal{B}_k L \hat{x}_{k|k}^r, \quad (22)$$

$$\mathcal{K}_k = \mathcal{P}_k \mathcal{C}_k^T \left(\mathcal{C}_k \mathcal{P}_k \mathcal{C}_k^T + \mathcal{R} \right)^{-1}, \quad (23)$$

$$\mathcal{P}_{k+1} = \mathcal{A}_k \left(\mathcal{P}_k - \mathcal{K}_k \mathcal{C}_k \mathcal{P}_k \right) \mathcal{A}_k^T + \mathcal{Q}. \quad (24)$$

Observe that we use the state estimate $\hat{x}_{k|k}^r$ to compute the input u_k^* as opposed to an estimate derived from (22). While this increases the cost J of control since an optimal state estimate is not used for state feedback, this prevents the attacker from using information from the input to learn about the system model. In fact, we have the following result.

Theorem 1: The input $u_k^* = L \hat{x}_{k|k}^r$ is independent from the system matrices $A_{1,k-1}, A_{2,k-1}, B_k, C_k$ for all k.

Proof: The input u_k^* is given by

$$l(\hat{x}_{0|0}^r, x_0, P, A, B, C, Q, R, w_0 \ldots w_{k-1}, v_1 \ldots v_k), \quad (25)$$

where l is some deterministic function of variables which by assumption are independent from $A_{1,k-1}, A_{2,k-1}, B_k, C_k$ for all k. As a result,

$$f_{u_k}(u_k | A_{1,0:k-1}, A_{2,0:k-1}, B_{0:k-1}, C_{1:k}) = f_{u_k}(u_k), \quad (26)$$

and the result holds. ∎

We assume that a residue based detector is incorporated where the residue z_k is given by

$$z_k \triangleq \begin{bmatrix} \tilde{y}_k \\ y_k \end{bmatrix} - \mathcal{C}_k \begin{bmatrix} \hat{\tilde{x}}_{k|k-1} \\ \hat{x}_{k|k-1} \end{bmatrix} \sim \mathcal{N}\left(0, \mathcal{C}_k \mathcal{P}_k \mathcal{C}_k^T + \mathcal{R}\right). \quad (27)$$

We can leverage knowledge of the distribution of z_k under normal operation to design a detector. In particular we consider χ^2 detector where g_k in (10) is given by

$$g_k(z_k) = z_k^T (\bar{\mathcal{P}}_k)^{-1} z_k, \quad (28)$$

where $\bar{\mathcal{P}}_k = \mathcal{C}_k \mathcal{P}_k \mathcal{C}_k + \mathcal{R}$. Under normal operation g_k has a χ^2 distribution. In general, the window for the detector can be extended to consider past measurements. In Figure 1, we include a diagram of the moving target system operating normally.
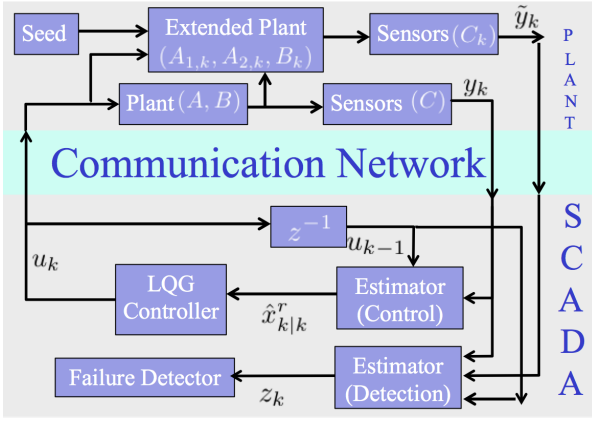
Fig. 1. Diagram of system under normal operation

IV. ATTACK MODEL

In this section we describe a near omniscient attacker in terms of his capabilities, access to information, and potential strategies. On one hand, the adversary may acquire his knowledge and resources through a highly sophisticated attack strategy as done in Stuxnet. On the other hand, an adversary can obtain his resources through insider information and access as done in the Maroochy Shire incident.

A. Attack Capabilities

1) The attacker can insert arbitrary inputs into the system and can arbitrarily alter the sensor measurements. As a result, when under attack, the system has dynamics given by

$$\begin{bmatrix} \tilde{x}_{k+1} \\ x_{k+1} \end{bmatrix} = \mathcal{A}_k \begin{bmatrix} \tilde{x}_k \\ x_k \end{bmatrix} + \mathcal{B}_k(u_k + u_k^a) + \begin{bmatrix} \tilde{w}_k \\ w_k \end{bmatrix}, \quad (29)$$

$$\begin{bmatrix} \tilde{y}_k^a \\ y_k^a \end{bmatrix} = \begin{bmatrix} \tilde{y}_k \\ y_k \end{bmatrix} + \begin{bmatrix} \tilde{s}_k^a \\ s_k^a \end{bmatrix}. \quad (30)$$

where u_k^a is the attacker's control input and $\tilde{s}_k^a$ and s_k^a are the biases injected on the extraneous sensors and ordinary sensors respectively.

2) The attacker can read the true outputs of the system $\tilde{y}_k, y_k$ and the inputs being sent by the defender to the plant u_k for all time k.

Remark 4: The attacker essentially performs a man in the middle attack between the plant and system operator so that he can manipulate and read all communication channels arbitrarily. A malicious insider can do this by breaking encryption schemes. Furthermore, physical attacks can be used to change sensor measurements. For instance, locally heating or cooling a temperature sensor would change the sensor measurements without violating the integrity or authenticity of data from a cyber perspective.

3) The attacker has full knowledge of the system model $\mathcal{S} \triangleq \{A, B, C, Q, R, K, L, \mathcal{Q}, \mathcal{R}\}$. Moreover the adversary knows the probability density function (pdf) of random matrices $A_{1,k}, A_{2,k}, B_k, C_{k+1}$.

Remark 5: While conservative, the adversary can obtain his knowledge of the system model by observing the communication channels for an extended period of time and

performing system identification. Moreover, observe that since the attacker is aware of the original system model and all outputs, he can asymptotically predict the state estimate $\hat{x}_{k|k}^r$ if the matrix $(A + BL)(I - KC)$ is stable [9].

Remark 6: The attacker can leverage his probabilistic knowledge of the system model as well as the true outputs of the system to generate stealthy attack inputs $s_k^a, \tilde{s}_k^a$. In particular, the adversary can attempt to simultaneously identify the moving target and generate convincing counterfeit sensor outputs.

Based on the above definitions we can define the private information available to the attacker and defender at time k $\mathcal{I}_k^A, \mathcal{I}_k^D$ and the public information $\mathcal{I}_k^P$ available to both as

$$\mathcal{I}_k^A \triangleq \{\tilde{y}_j, y_j, \tilde{s}_j^a, s_j^a, u_{j-1}^a\} \quad \forall j \leq k, \quad (31)$$

$$\mathcal{I}_k^D \triangleq \{A_{1,j-1}, A_{2,j-1}, B_{j-1}, C_j\} \quad \forall j, \quad (32)$$

$$\mathcal{I}_k^P \triangleq \{\mathcal{S}, f(A_1, A_2, B, C), u_{j-1}, \tilde{y}_j^a, y_j^a\} \quad \forall j \leq k. \quad (33)$$

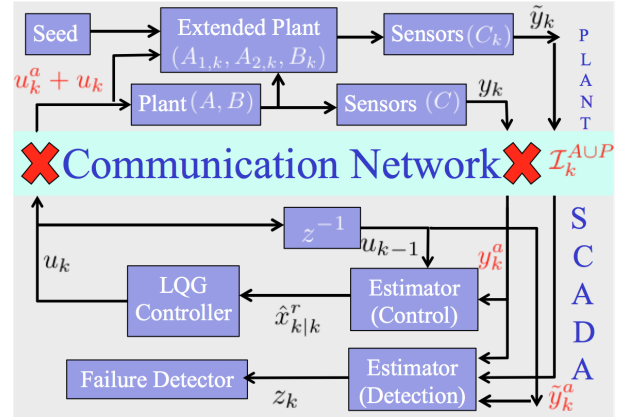In Figure 2, we include a diagram of the system under attack.

Fig. 2. Diagram of system under attack

B. Attack Strategy

In this subsection we propose two main attack strategies. Without loss of generality we assume any injection attack begins at $k = 0$.

1) Attack 1: Subtract Influence: In the first attack strategy the attacker aims to estimate his influence on the control system and subtract it. Define $\bar{x}_k \triangleq [\tilde{x}_k^T \ x_k^T]^T$, $\bar{y}_k \triangleq [\tilde{y}_k^T \ y_k^T]^T$, $\bar{s}_k^a \triangleq [\tilde{s}_k^a{}^T \ s_k^a{}^T]^T$. We observe that the attacker's influence on the system is given by the output $\bar{y}_k^a$ of the following linear time-varying system.

$$\bar{x}_{k+1}^a = \mathcal{A}_k \bar{x}_k^a + \mathcal{B}_k u_k^a, \quad \Delta \bar{y}_k^a = \mathcal{C}_k \bar{x}_k^a, \quad (34)$$

with initial state $\bar{x}_0^a = 0$. In this attack, the adversary aims to compute an estimate of $\Delta \bar{y}_k^a$, ideally obtaining

$$\bar{s}_k^a = -\mathbb{E}[\Delta \bar{y}_k^a | \mathcal{I}_k^a \cup \mathcal{I}_k^P]. \quad (35)$$

Remark 7: Observe that the adversary can exactly subtract his influence from measurements y_k due to his knowledge of the system model. However, the adversary should be unable to subtract his bias from the extraneous sensors $\tilde{y}_k$.

Optimal Theoretical Estimation Define $\bar{w}_k \triangleq [\tilde{w}_k^T \ w_k^T]^T$, $\bar{v}_k \triangleq [\tilde{v}_k^T \ v_k^T]^T$, and $\bar{y}_k \triangleq [\tilde{y}_k^{aT} \ y_k^{aT}]^T$. The adversary's observations can be formulated through the following linear time-varying system,

$$\begin{bmatrix} \bar{x}_{k+1} \\ \bar{x}_{k+1}^a \end{bmatrix} = \begin{bmatrix} \mathcal{A}_k & 0 \\ 0 & \mathcal{A}_k \end{bmatrix} \begin{bmatrix} \bar{x}_k \\ \bar{x}_k^a \end{bmatrix} + \begin{bmatrix} \mathcal{B}_k & \mathcal{B}_k \\ 0 & \mathcal{B}_k \end{bmatrix} \begin{bmatrix} u_k \\ u_k^a \end{bmatrix} + \begin{bmatrix} \bar{w}_k \\ 0 \end{bmatrix}, \tag{36}$$

$$\bar{y}_k = \begin{bmatrix} \mathcal{C}_k & 0 \end{bmatrix} \begin{bmatrix} \bar{x}_k \\ \bar{x}_k^a \end{bmatrix} + \bar{v}_k. \tag{37}$$

To estimate $\Delta \bar{y}_k^a$ at time k, assume the adversary has access to the following distribution $f(\bar{x}_k, \bar{x}_k^a, C_k | \mathcal{I}_k^{A\cup P})$ where $\mathcal{I}_k^{A\cup P} = \mathcal{I}_k^A \cup \mathcal{I}_k^P$ Then we have

$$\bar{s}_k^a = -\int_{\bar{x}_k} \int_{\bar{x}_k^a} \int_{C_k} \mathcal{C}_k \bar{x}_k^a f(\bar{x}_k, \bar{x}_k^a, C_k | \mathcal{I}_k^{A\cup P}) \mathrm{d}\bar{x}_k \mathrm{d}\bar{x}_k^a \mathrm{d}C_k. \tag{38}$$

We show that the pdf can be recursively computed at each step. Letting $\zeta_{k+1} = \{\bar{x}_{k+1}, \bar{x}_{k+1}^a, C_{k+1}\}$ we have

$$\begin{aligned} f(\zeta_{k+1} | \mathcal{I}_{k+1}^{A\cup P}) &= f(\zeta_{k+1} | \mathcal{I}_k^{A\cup P}, \bar{y}_{k+1}^a, \bar{y}_{k+1}, \bar{s}_{k+1}^a, u_k^a, u_k), \\ &= f(\zeta_{k+1} | \mathcal{I}_k^{A\cup P}, \bar{y}_{k+1}, u_k^a, u_k), \\ &= \frac{f(\bar{y}_{k+1} | \mathcal{I}_k^{A\cup P}, \zeta_{k+1}) f(\zeta_{k+1} | \mathcal{I}_k^{A\cup P}, u_k, u_k^a)}{f(\bar{y}_{k+1} | \mathcal{I}_k^{A\cup P}, u_k, u_k^a)}. \end{aligned} \tag{39}$$

The second equality follows from the conditional independence of ζ_{k+1} and $\bar{y}_{k+1}^a, \bar{s}_{k+1}^a$ given $\bar{y}_{k+1}$. The last equality follows from Bayes rule and the conditional independence of $\bar{y}_{k+1}$ and u_k, u_k^a given ζ_{k+1}. We note that this distribution can be theoretically computed given the attacker's information. That is we know that,

$$f(\bar{y}_{k+1} | \mathcal{I}_k^{A\cup P}, \zeta_{k+1}) \sim \mathcal{N}(\mathcal{C}_{k+1} \bar{x}_{k+1}, \mathcal{R}). \tag{40}$$

Moreover, ζ_{k+1} and $\bar{y}_{k+1}$ are deterministic functions of ζ_k, u_k, u_k^a and random variables $A_{1,k}$, $A_{2,k}$, B_k, C_{k+1}, $\bar{w}_k$, $\bar{v}_{k+1}$ which are independent of $\mathcal{I}_k^{A\cup P}$. Thus, the associated density functions can be computed from $f(\zeta_k | \mathcal{I}_k^{A\cup P})$.

Remark 8: If the attacker subtracts his influence, he might be susceptible to a growing cancellation error if he attempts to excite the system's unstable dynamics. Instead of subtracting his influence the attacker can instead directly estimate what the defender expects to see as summarized in the next section.

2) Attack 2: Estimate Expected Measurement: In the next strategy, the adversary aims to track the system operator's state estimate. Using the system operator's state estimate, the adversary attempts to generate stealthy outputs. Let $\hat{\bar{x}}_k = [\hat{\bar{x}}_{k|k-1}^T \hat{x}_{k|k-1}^T]^T$. The attacker's observations and strategy can be formulated as follows

$$\begin{bmatrix} \bar{x}_{k+1} \\ \hat{\bar{x}}_{k+1} \end{bmatrix} = \begin{bmatrix} \mathcal{A}_k & 0 \\ 0 & \mathcal{A}_k(I - \mathcal{K}_k \mathcal{C}_k) \end{bmatrix} \begin{bmatrix} \bar{x}_k \\ \hat{\bar{x}}_k \end{bmatrix} + \begin{bmatrix} \bar{w}_k \\ 0 \end{bmatrix},$$

$$+ \begin{bmatrix} \mathcal{B}_k & \mathcal{B}_k & 0 \\ \mathcal{B}_k & 0 & \mathcal{A}_k \mathcal{K}_k \end{bmatrix} \begin{bmatrix} u_k \\ u_k^a \\ \bar{y}_k^a \end{bmatrix}, \tag{41}$$

$$\bar{y}_k = \begin{bmatrix} \mathcal{C}_k & 0 \end{bmatrix} \begin{bmatrix} \bar{x}_k \\ \hat{\bar{x}}_k \end{bmatrix} + \bar{v}_k, \quad \bar{s}_k^a = \mathbb{E}[\mathcal{C}_k \hat{\bar{x}}_k | \mathcal{I}_k^{A\cup P}] - \bar{y}_k. \tag{42}$$

The use of the preceding attack design is motivated by the ensuing result which states that the chosen attack vector minimizes the expected value of our threshold detector.

Theorem 2: Consider (10) with g_k given by (28). Then

$$\mathbb{E}[\mathcal{C}_k \hat{\bar{x}}_k | \mathcal{I}_k^{A\cup P}] - \bar{y}_k = \arg\min_{\bar{s}_k^a} \mathbb{E}[g_k | \mathcal{I}_k^{A\cup P}]. \tag{43}$$

Proof: Let $\zeta_k = \{\bar{x}_k, \hat{\bar{x}}_k, C_k, \mathcal{P}_k\}$. Observe that

$$\mathbb{E}[g_k | \mathcal{I}_k^{A\cup P}] = \int_{\zeta_k} z_k^T (\mathcal{C}_k \mathcal{P}_k \mathcal{C}_k^T + \mathcal{R})^{-1} z_k f(\zeta_k | \mathcal{I}_k^{A\cup P}) \mathrm{d}\zeta_k. \tag{44}$$

Taking the gradient with respect to $\bar{s}_k^a$ and setting the equation equal to 0 gives

$$\int_{\zeta_k} (\mathcal{C}_k \mathcal{P}_k \mathcal{C}_k^T + \mathcal{R})^{-1} (\bar{y}_k + s_k^a - \mathcal{C}_k \hat{\bar{x}}_k) f(\zeta_k | \mathcal{I}_k^{A\cup P}) \mathrm{d}\zeta_k = 0. \tag{45}$$

Solving gives

$$\bar{s}_k^a = -\bar{y}_k + \int_{\zeta_k} \mathcal{C}_k \hat{\bar{x}}_k f(\zeta_k | \mathcal{I}_k^{A\cup P}) \mathrm{d}\zeta_k, \tag{46}$$

and the result holds. ∎

To determine $\bar{s}_k^a$ at time k assume the adversary has access to the following distribution $f(\zeta_k | \mathcal{I}_k^{A\cup P})$. As done before, the attacker can theoretically compute $\bar{s}_k^a$ by taking the conditional expectation. Additionally, we can show that (39) holds for the case of attack 2. Moreover, by similar analysis as in attack 1, we can demonstrate the density functions in (39) can theoretically be computed. The main difference here is that the adversary must also estimate $\mathcal{P}_k$. Note that in practice the proposed attacks are difficult to execute for an adversary since it is likely a challenge to compute the necessary distribution functions and expected values. As a result, in the next section we aim to provide bounds on the attacker's estimation performance in terms of mean square error matrices.

V. BOUNDS ON ATTACKER'S PERFORMANCE

A. Bounds on Attacker's State Estimation

In this section we attempt to characterize lower bounds on the error matrices associated with the states ζ_k defined in attack strategy 1 and 2. From there, we can attempt to characterize how well the adversary can design $\bar{s}_k^a$ to fool the bad data detector.

We leverage conditional posterior Cramer-Rao lower bounds for Bayesian sequences derived by [17]. The authors here make use of the Bayesian Cramer-Rao lower bound or Van Trees bound derived in [18] which states that for observations y and states ζ the mean squared error matrix is bounded by the Fisher information as follows

$$\mathbb{E}_{f(\zeta,y)} \left[[\hat{\zeta}(y) - \zeta][\hat{\zeta}(y) - \zeta]^T \right] \geq I^{-1}, \tag{47}$$

where the Fisher information matrix I is given by

$$I = \mathbb{E}_{f(\zeta,y)} \left[-\triangle_\zeta^\zeta \log f(\zeta, y) \right]. \tag{48}$$

Note that

$$\triangle_x^y g(x, y) \triangleq \nabla_x \nabla_y^T g(x, y).$$

In [17], this result is extended to nonlinear Bayesian sequences with dynamics given by

$$\zeta_{k+1} = F_k(\zeta_k, \omega_k), \quad y_k = G_k(\zeta_k, v_k), \qquad (49)$$

where ω_k and v_k are independent process and sensor noise respectively. In our case, we slightly adapt these results to account for the fact there is feedback in our system so that

$$\zeta_{k+1} = F_k(\zeta_k, \bar{y}_{1:k}, \omega_k), \quad \bar{y}_k = G_k(\zeta_k, v_k). \qquad (50)$$

The inputs u_k, u_k^a and $\bar{s}_k^a$ are incorporated into the definition of F_k, while uncertainty in the model $(A_{1,k}, A_{2,k}, B_k, C_{k+1})$ can be incorporated in the process noise ω_k. It can shown that the following posterior Cramer-Rao lower bound holds

$$\mathbb{E}_{f_{k+1}^c}\left[e_{0:k+1}e_{0:k+1}^T|\bar{y}_{1:k}\right] \geq I^{-1}(\zeta_{0:k+1}|\bar{y}_{1:k}), \qquad (51)$$

where

$$e_{0:k+1} \triangleq \zeta_{0:k+1} - \hat{\zeta}_{0:k+1}(\bar{y}_{k+1}|\bar{y}_{1:k}), \qquad (52)$$

$$f_{k+1}^c \triangleq f(\zeta_{0:k+1}, \bar{y}_{k+1}|\bar{y}_{1:k}), \qquad (53)$$

$$I(\zeta_{0:k+1}|\bar{y}_{1:k}) \triangleq \mathbb{E}_{f_{k+1}^c}\left[-\triangle_{\zeta_{0:k+1}}^{\zeta_{0:k+1}}\log f_{k+1}^c|\bar{y}_{1:k}\right]. \qquad (54)$$

We observe that (51) gives us an expected lower bound for the error matrix associated with the entire state history $\zeta_{0:k+1}$ with knowledge of measurements $\bar{y}_{1:k}$. This expectation is taken over the state history as well the measurement $\bar{y}_{k+1}$ so that $\hat{\zeta}_{0:k+1}$ is a function of the measurement $\bar{y}_{k+1}$. Observe that unlike the traditional Cramer-Rao bound which is limited to unbiased estimators, the Bayesian Cramer-Rao bound here considers both biased and unbiased estimators $\hat{\zeta}$.

While the lower bound given here applies to the entire state history $\zeta_{0:k+1}$, in practice we care about estimating a lower bound on the current state ζ_{k+1}. Nonetheless, it can be easily shown that

$$\mathbb{E}_{f_{k+1}^c}\left[e_{k+1}e_{k+1}^T|\bar{y}_{1:k}\right] \geq I^{-1}(\zeta_{k+1}|\bar{y}_{1:k}), \qquad (55)$$

where $I^{-1}(\zeta_{k+1}|\bar{y}_{1:k})$ is the $\dim(\zeta_k) \times \dim(\zeta_k)$ lower right submatrix of $I^{-1}(\zeta_{0:k+1}|\bar{y}_{1:k})$. In practice, computing $I^{-1}(\zeta_{k+1}|\bar{y}_{1:k})$ from $I^{-1}(\zeta_{0:k+1}|\bar{y}_{1:k})$ is impractical since it requires computing and taking the inverse of a Fisher information matrix which grows in dimension at each time step. As a result, we would like a recursion to compute $I^{-1}(\zeta_{k+1}|\bar{y}_{1:k})$. From [17] we have the following result,

$$I(\zeta_{k+1}|\bar{y}_{1:k}) = D_k^{22} - D_k^{21}\left[D_k^{11} + I_A(\zeta_k|\bar{y}_{1:k})\right]^{-1}D_k^{12}, \qquad (56)$$

where

$$D_k^{11} = \mathbb{E}_{f_{k+1}^c}\left[-\triangle_{\zeta_k}^{\zeta_k}\log f(\zeta_{k+1}|\zeta_k, \bar{y}_{1:k})\right],$$

$$D_k^{12} = \mathbb{E}_{f_{k+1}^c}\left[-\triangle_{\zeta_k}^{\zeta_{k+1}}\log f(\zeta_{k+1}|\zeta_k, \bar{y}_{1:k})\right] = (D_k^{21})^T,$$

$$D_k^{22} = \mathbb{E}_{f_{k+1}^c}\left[-\triangle_{\zeta_{k+1}}^{\zeta_{k+1}}\log f(\zeta_{k+1}|\zeta_k, \bar{y}_{1:k})f(\bar{y}_{k+1}|\zeta_{k+1})\right].$$

In addition,

$$I_A(\zeta_k|\bar{y}_{1:k}) = E_k^{22} - E_k^{21}\left(E_k^{11}\right)^{-1}E_k^{12}, \qquad (57)$$

where

$$E_k^{11} = \mathbb{E}_{f(\zeta_{0:k}|\bar{y}_{1:k})}\left[-\triangle_{\zeta_{0:k-1}}^{\zeta_{0:k-1}}\log f(\zeta_{0:k}|\bar{y}_{1:k})\right],$$

$$E_k^{12} = \mathbb{E}_{f(\zeta_{0:k}|\bar{y}_{1:k})}\left[-\triangle_{\zeta_{0:k-1}}^{\zeta_k}\log f(\zeta_{0:k}|\bar{y}_{1:k})\right] = (E_k^{21})^T,$$

$$E_k^{22} = \mathbb{E}_{f(\zeta_{0:k}|\bar{y}_{1:k})}\left[-\triangle_{\zeta_k}^{\zeta_k}\log f(\zeta_{0:k}|\bar{y}_{1:k})\right].$$

We observe that it is still difficult to obtain matrices $E_k^{11}, E_k^{12}, E_k^{21}, E_k^{22}$ so [17] introduces the following approximate recursion

$$I_A(\zeta_k|\bar{y}_{1:k}) \approx S_k^{22} - S_k^{12\ T}\left[S_k^{11} + I_A(\zeta_{k-1}|\bar{y}_{1:k-1})\right]^{-1}S_k^{12}, \qquad (58)$$

where

$$S_k^{11} = \mathbb{E}_{f(\zeta_{0:k}|\bar{y}_{1:k})}\left[-\triangle_{\zeta_{k-1}}^{\zeta_{k-1}}\log f(\zeta_k|\zeta_{k-1}, \bar{y}_{1:k-1})\right],$$

$$S_k^{12} = \mathbb{E}_{f(\zeta_{0:k}|\bar{y}_{1:k})}\left[-\triangle_{\zeta_{k-1}}^{\zeta_k}\log f(\zeta_k|\zeta_{k-1}, \bar{y}_{1:k-1})\right],$$

$$S_k^{22} = \mathbb{E}_{f(\zeta_{0:k}|\bar{y}_{1:k})}\left[-\triangle_{\zeta_k}^{\zeta_k}\log f(\zeta_k|\zeta_{k-1}, \bar{y}_{1:k-1})f(\bar{y}_k|\zeta_k)\right].$$

We observe that in practice it may still be difficult to compute the exact expectations because high dimensional integration is generally involved. Nonetheless, particle filters as described in [19] can be used to approximate these expectations. Alternative approximations for the conditional posterior Cramer-Rao lower bound can be found in [20]. Unconditional bounds can be found in [21].

B. Bounds on Detection

The algorithm described allows us to compute an approximate lower bound on the mean square error matrix of the attacker's state ζ_k for a given set of inputs $u_k^a, \bar{s}_k^a$ and observation history $\bar{y}_{1:k}$. In the case of attack strategy 2, this allows us to obtain a lower bound on the value of $g_k(z_k)$ as follows.

Theorem 3: Consider attack strategy 2. For time $k + 1$, redefine F_k and f_{k+1}^c so C_{k+1} is known to the adversary (corresponding to reducing the variance of the process noise ω_k). Let $\hat{\tilde{x}}_{k+1}^e(\bar{y}_{k+1})$ be an estimate of $\hat{\tilde{x}}_{k+1}$ given $\bar{y}_{1:k}$ and $\hat{e}_{k+1} = \hat{\tilde{x}}_{k+1} - \hat{\tilde{x}}_{k+1}^e(\bar{y}_{k+1})$. Suppose a lower bound Z on the error matrix of $\hat{\tilde{x}}_{k+1}$ is obtained so that

$$\mathbb{E}_{f_{k+1}^c}\left[\hat{e}_{k+1}\hat{e}_{k+1}^T\right] \geq Z. \qquad (59)$$

Then we have

$$\min_{\hat{\tilde{x}}_{k+1}^e} \mathbb{E}_{f^*}\left[g_{k+1}(z_{k+1})\right] \geq \text{tr}(\mathcal{C}_{k+1}Z\mathcal{C}_{k+1}^T\bar{\mathcal{P}}_{k+1}^{-1}), \qquad (60)$$

where $f^* = f(\hat{\tilde{x}}_{k+1}, C_{k+1}, \mathcal{I}_{k+1}^{A\cup P}/\mathcal{I}_k^{A\cup P}|\mathcal{I}_k^{A\cup P})$.

Proof: Let $\mathcal{I}_{k+1}^{*\ A\cup P} = \mathcal{I}_{k+1}^{A\cup P}/\mathcal{I}_k^{A\cup P}$. We observe the following result.

$$\min_{\hat{\tilde{x}}_{k+1}^e} \mathbb{E}_{f^*}[g_{k+1}] \geq \min_{\hat{\tilde{x}}_{k+1}^e} \mathbb{E}_{f(\hat{\tilde{x}}_{k+1}, \mathcal{I}_{k+1}^{*\ A\cup P}|\mathcal{I}_k^{A\cup P}, C_{k+1})}[g_{k+1}],$$

$$= \min_{\hat{\tilde{x}}_{k+1}^e} \mathbb{E}_{f_{k+1}^c}[z_{k+1}^T\bar{\mathcal{P}}_{k+1}^{-1}z_{k+1}],$$

$$= \min_{\hat{\tilde{x}}_{k+1}^e} \text{tr}(\mathbb{E}_{f_{k+1}^c}[z_{k+1}z_{k+1}^T]\bar{\mathcal{P}}_{k+1}^{-1}), \qquad (61)$$

$$= \text{tr}(\mathcal{C}_{k+1}\mathbb{E}_{f_{k+1}^c}[\hat{e}_{k+1}^*\hat{e}_{k+1}^{*T}]\mathcal{C}_{k+1}^T\bar{\mathcal{P}}_{k+1}^{-1}),$$

$$\geq \text{tr}(\mathcal{C}_{k+1}Z\mathcal{C}_{k+1}^T\bar{\mathcal{P}}_{k+1}^{-1}), \qquad (62)$$

where $\hat{e}^*_{k+1}$ is the minimizer of (61). The first inequality follows from the monotonicity of g_{k+1} in the error matrix of $\mathcal{C}_{k+1}\hat{\bar{x}}_{k+1}$ and the fact that the error matrix is reduced by providing more information, C_{k+1}. The first equality follows from the definition of f^c_{k+1} which is a density function conditioned on $\bar{y}_{1:k}$. Additional information in $\mathcal{I}^{A\cup P}_k$ is included in f^c_{k+1} through the definitions of F_k and G_k. The second and third equalities follow from the linearity of the expectation. The final inequality follows from (59). ∎

Remark 9: In general, the adversary's ability to estimate ζ_k is dependent on the inputs $u^a_k, \bar{s}^a_k$. For instance, the more the adversary biases the state away from its expected region of operation, the more challenging it is to perform estimation. Thus, if the system operator wishes to analyze how well an adversary can generate stealthy outputs, he must consider a particular sequence of attack inputs $u^a_k, \bar{s}^a_k$.

Remark 10: In practice, it may be difficult to perform performance analysis when assuming $\mathcal{P}_k$ is an unknown state. However, one can still obtain a lower bound on the error matrix by assuming that the adversary has an oracle which allows him to know $\mathcal{P}_k$, $\mathcal{K}_k$, $I - \mathcal{K}_k\mathcal{C}_k$. This likely makes the computational analysis of attack strategy 2 easier.

VI. CONCLUSION

In this paper, we have considered attacks on control systems where an adversary has access to all channels in a communication network. In order to counter such an adversary, we propose introducing time-varying dynamics into the system which are unknown to the adversary and can in turn be leveraged to detect attacks. Future work will consider sufficient conditions for the design of these matrices to prevent zero-dynamic attacks and the analysis of optimal identification techniques for the adversary.

REFERENCES

[1] T. M. Chen, "Stuxnet, the real start of cyber warfare? [editor's note]," *IEEE Network*, vol. 24, no. 6, pp. 2–3, 2010.

[2] R. Langner, "To kill a centrifuge: A technical analysis of what Stuxnet's creators tried to achieve," Langner Communications, Tech. Rep., November 2013. [Online]. Available: www.langner.com/en/wp-content/uploads/2013/11/To-kill-a-centrifuge.pdf

[3] J. Slay and M. Miller, "Lessons learned from the Maroochy water breach," in *Critical Infrastructure Protection*. Springer US, 2008, pp. 73–82.

[4] A. A. Cárdenas, S. Amin, and S. S. Sastry, "Secure Control: Towards Survivable Cyber-Physical Systems," in *Distributed Computing Systems Workshops, 2008. ICDCS '08. 28th International Conference on DOI - 10.1109/ICDCS.Workshops.2008.40*. IEEE, 2008, pp. 495–500.

[5] A. Teixeira, D. Perez, H. Sandberg, and K. H. Johannson, "Attack models and scenarios for networked control systems," in *Proceedings of the 1st international conference on High Confidence Networked Systems*, Beijing, China, 2012, pp. 55–64.

[6] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[7] Y. Liu, M. Reiter, and P. Ning, "False data injection attacks against state estimation in electric power grids," in *Proceedings of the 16th ACM conference on computer and communications security*, Chicago, IL, 2009.

[8] Y. Mo and B. Sinopoli, "False data injection attacks in cyber physical systems," in *First Workshop on Secure Control Systems*, Stockholm, Sweden, April 2010.

[9] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on SCADA systems," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, 2014.

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

[11] H. Sandberg, A. Teixeira, and K. H. Johansson, "On security indices for state estimators in power networks," in *First Workshop on Secure Control Systems*, Stockholm, Sweden, 2010.

[12] S. Weerakkody, Y. Mo, and B. Sinopoli, "Detecting integrity attacks on control systems using robust physical watermarking," in *53rd IEEE Conference on Decision and Control*, Los Angeles, California, 2014, pp. 3757–3764.

[13] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1454–1467, 2014.

[14] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "Revealing stealthy attacks in control systems," in *50th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, Illinois, 2012, pp. 1806–1813.

[15] P. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Prentice Hall, 1986.

[16] R. Smith, "A decoupled feedback structure for covertly appropriating network control systems," in *IFAC World Congress*, Milan, Italy, 2011, pp. 90–95.

[17] L. Zuo, R. Niu, and P. K. Varshney, "Conditional posterior Cramer - Rao lower bounds for nonlinear sequential Bayesian estimation," *IEEE Transactions on Signal Processing*, vol. 59, no. 1, pp. 1–14, 2011.

[18] H. L. Van Trees, *Detection Estimation and Modulation Theory*. New York: Wiley, 1968, vol. 1.

[19] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[20] Y. Zheng, O. Ozdemir, R. Niu, and P. K. Varshney, "New conditional posterior Cramer - Rao lower bounds for nonlinear sequential Bayesian estimation," *IEEE Transactions on Signal Processing*, vol. 60, no. 10, pp. 5549–5556, 2012.

[21] P. Tichavsky, C. H. Muravchik, and A. Nehorai, "Posterior Cramer - Rao bounds for discrete-time nonlinear filtering," *IEEE Transactions on Signal Processing*, vol. 48, no. 2, pp. 1386–1395, 1998.