On the Confidentiality of Linear Anomaly Detector States

David Umsonst, Ehsan Nekouei, André Teixeira, Henrik Sandberg

***Abstract*— A malicious attacker with access to the sensor channel in a feedback control system can severely affect the physical system under control, while simultaneously being hard to detect. A properly designed anomaly detector can restrict the impact of such attacks, however. Anomaly detectors with an internal state (stateful detectors) have gained popularity because they seem to be able to mitigate these attacks more than detectors without a state (stateless detectors). In the analysis of attacks against control systems with anomaly detectors, it has been assumed that the attacker has access to the detector's internal state, or designs its attack such that it is not detected regardless of the detector's state. In this paper, we show how an attacker can realize the first case by breaking the confidentiality of a stateful detector state evolving with linear dynamics, while remaining undetected and imitating the statistics of the detector under nominal conditions. The realization of the attack is posed in a convex optimization framework using the notion of Kullback-Leibler divergence. Further, the attack is designed such that the maximum mean estimation error of the Kalman filter is maximized at each time step by exploiting dual norms. A numerical example is given to illustrate the results.**

I. INTRODUCTION

Utilizing communication networks to reduce the cost and increase the efficiency in control systems has created so called cyber-physical systems (CPSs). CPS are not limited to industrial processes but include also critical infrastructures such as the power grid and water distribution grids. Due to the communication networks CPSs are faced with the threat of cyber-attacks.

Therefore, an investigation in control-theoretic methods to enhance the security of CPS has begun in recent years. These control-theoretic methods can be seen as an additional layer to the information technology related security measures such as cryptography and authentication. Using watermarking of control or sensor signals to improve the security of CPS is a common approach. Watermarking is both used as an additive [1] and a multiplicative [2] signal. Furthermore, Hespanhol *et al.* [3] use watermarking in networked control systems to detect the presence of attacks. Another research direction deals with the estimation of the possible attack impact. Milošević *et al.* [4] proposes a unifying framework for several attack strategies and estimating their impact on discrete-time deterministic linear systems, while [5] defines the identifiability and detectability of attacks on CPS.

Anomaly detectors have recently gained more interest in the research community. Typically detectors are used to detect randomly occurring faults in a system, but their abilities to detect attacks or mitigate the impact of attacks that avoid detection become more important. Investigating new anomaly detectors such as hybrid detectors [6] is of equal importance as determining the performance of commonly used anomaly detector under attack, where our work focuses on the latter. Commonly used anomaly detectors that have been investigated recently are the χ^2, cumulative sum (CUSUM) [7], and the multivariate exponentially weighted moving average (MEWMA) [8] detector. Detectors such as the CUSUM and MEWMA detector have internal states, which seems to benefit the attack impact mitigation. A metric to compare the detector performance in the presence of attacks is introduced by Urbina *et al.* [9] and based on the mean time between false alarms and the impact of an undetectable attack on the system. Murguia *et al.* [10] investigate full sensor attacks on control systems equipped with χ^2 or CUSUM detectors and propose tuning methods for these detectors. In our previous work [11] we investigate the impact of full sensor attacks on control systems with a general anomaly detector model, which includes the χ^2, CUSUM, and MEWMA detector. Furthermore, we propose an extension of the metric in [9] for the case of full sensor attacks in [11], where the attacker has no knowledge of the internal state of the detector.

Often it is assumed that the attacker has knowledge of the detector's state to determine a worst-case attack impact, e.g. in [9], [10], or the attacker designs its attack such that it is not detected irrespectively of the detector's state [11]. Therefore we investigate the confidentiality of the internal state of an anomaly detector in this paper.

Our contribution is threefold. First, we show how an undetectable attack is able to break the confidentiality of the internal state of a detector that evolves with linear dynamics, which include for example the MEWMA detector and also the generalized MEWMA detector [12]. Second, not only is the confidentiality broken, but the attacker is also able to imitate the statistics of the detector under nominal condition by utilizing the Kullback-Leibler divergence. Last, we show how the attack can be designed, such that the mean of the Kalman filter's estimation error is maximized at each time step. This method of breaking the detector state's confidentiality is verified for a MEWMA detector.

The rest of the paper has the following structure. The CPS, anomaly detector, and attacker model used in our work is presented in Section II. The method how an attacker can break

This work was supported in part by the Swedish Research Council (grants 2016-00861 and 2018-04396), the Swedish Civil Contingencies Agency through the CERCES project, and the Swedish Energy Agency through the ERA-Net project LarGo!.

David Umsonst, Ehsan Nekouei, and Henrik Sandberg are with the Division of Decision and Control Systems in the School of Electrical Engineering and Computer Science at the KTH Royal Institute of Technology, 10044 Stockholm, Sweden {umsonst,nekouei,hsan}@kth.se

André Teixeira is with the Department of Engineering Sciences, Signals and Systems Group at Uppsala University, Sweden. andre.teixeira@angstrom.uu.se

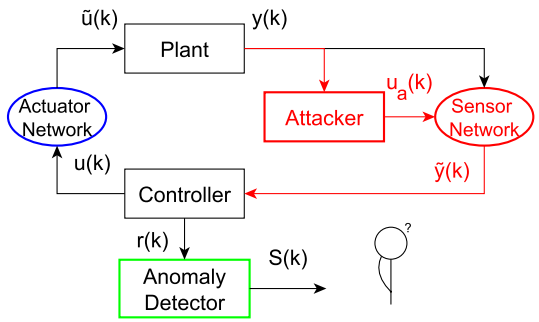

Fig. 1. Block Diagram of the Attack Scenario

the confidentiality is presented in Section III and applied to the MEWMA detector in Section IV. To verify our results a numerical example for the MEWMA detector is given in Section IV as well. Our work concludes in Section V with an outlook on possible future directions.

II. SYSTEM MODEL

In this section, we explain the attack scenario considered in this paper. In this scenario, a plant is controlled over wireless networks and the controller side of the system is equipped with an anomaly detector. An attacker managed to penetrate the sensor network. Figure 1 shows the block diagram of the attack scenario, which we describe in detail in this section.

A. Plant and Controller Model

Due to the connection with a wireless network we model the plant and the controller as linear discrete-time systems. The plant dynamics are given by

$$x(k+1) = Ax(k) + B\tilde{u}(k) + w(k) \tag{1}$$
$$y(k) = Cx(k) + v(k), \tag{2}$$

where $x(k) \in \mathbb{R}^{n_x}$ is the state of the plant, $\tilde{u}(k) \in \mathbb{R}^{n_u}$ is the control input received over the network and $y(k) \in \mathbb{R}^{n_y}$ is the measurement signal of the plant, all at time step $k \in \mathbb{N}_{\geq 0}$. The system matrix is given by $A \in \mathbb{R}^{n_x \times n_x}$, while $B \in \mathbb{R}^{n_x \times n_u}$, and $C \in \mathbb{R}^{n_y \times n_x}$ describe the influence of the input on the state and define the measurements taken, respectively. The states of the plant have an additive process noise $w(k) \sim \mathcal{N}(0, \Sigma_w)$, while the measurements have an additive measurement noise $v(k) \sim \mathcal{N}(0, \Sigma_v)$. The initial state is $x(0) \sim \mathcal{N}(0, \Sigma_x)$. Here, $\Sigma_w \in \mathbb{R}^{n_x \times n_x}$, $\Sigma_v \in \mathbb{R}^{n_y \times n_y}$, and $\Sigma_x \in \mathbb{R}^{n_x \times n_x}$ are positive definite covariance matrices, respectively, and these white Gaussian processes are mutually independent. The controller is designed as a linear quadratic Gaussian controller

$$\hat{x}(k+1) = A\hat{x}(k) + Bu(k) + L(k)(\tilde{y}(k) - C\hat{x}(k)) \tag{3}$$
$$u(k) = -K\hat{x}(k), \tag{4}$$
$$\tilde{r}(k) = \tilde{y}(k) - C\hat{x}(k), \tag{5}$$

where $\hat{x}(k) \in \mathbb{R}^{n_x}$ is the Kalman filter's estimate of $x(k)$, $\hat{x}(0) = 0$, $u(k) \in \mathbb{R}^{n_u}$ is the actuator signal determined by the controller, $L(k)$ is the Kalman gain, K is the controller gain, which is designed such that $\rho(A - BK) < 1$, where $\rho(A)$ is the spectral radius of matrix A, $\tilde{r}(k)$ is the residual, which is the difference between the received measurements $\tilde{y}(k) \in \mathbb{R}^{n_y}$ over the network and the estimated system output $C\hat{x}(k)$. In the case of no attack, the residual is an independent Gaussian random variable with $\tilde{r}(k) \sim \mathcal{N}(0, \Sigma_r(k))$, where $\Sigma_r(k) \in \mathbb{R}^{n_y \times n_y}$ is a positive definite covariance matrix.

Assumption 1: The system has reached steady state before the attack happens, such that $L(k) = L$ and $\Sigma_r(k) = \Sigma_r$, where $L = \lim_{k\to\infty} L(k)$ and $\Sigma_r = \lim_{k\to\infty} \Sigma_r(k)$ are the steady state values of the Kalman gain and the residual covariance matrix, respectively. For the steady steady state values to exists, we further assume (A, C) is detectable and $(A, \Sigma_w^{\frac{1}{2}})$ is controllable. Further, the network works perfectly except for the attack. Hence, $\tilde{u}(k) = u(k)$, while $\tilde{y}(k) \neq y(k)$ due to the attack.

With this assumption we define $r(k) = \Sigma_r^{-\frac{1}{2}} \tilde{r}(k)$ as the normalized residual signal, where $r(k) \sim \mathcal{N}(0, I)$ in the nominal case, i.e. when $y(k) = \tilde{y}(k)$.

B. Anomaly Detectors

Anomaly detectors in control systems are typically used to detect randomly occuring faults in the plant [13]. The detector can be described by a possibly nonlinear discrete-time system

$$\begin{aligned} x_D(k+1) &= \theta\big(x_D(k), r(k)\big), \\ S(k+1) &= d\big(x_D(k), r(k)\big), \end{aligned} \tag{6}$$

where $x_D(k) \in \mathbb{R}^{n_D}$ is the internal state of the detector, which is initialized as a zero vector, $S(k+1) \in \mathbb{R}_{\geq 0}$ can be seen as the output of the detector, and $r(k)$ is the input to the detector. Here, $\theta\big(x_D(k), r(k)\big)$ describes the dynamics of the detector state, and $d\big(x_D(k), r(k)\big)$ the output behaviour of the detector. If the detector has no internal state, we call it *stateless*, and *stateful* otherwise.

A small $S(k+1)$, i.e. $S(k+1) \approx 0$, indicates that no anomalies are present. If $S(k+1) > J_D$, where $J_D \geq 0$, an alarm is triggered. In case no fault or intruder is present, we call it a *false alarm*. To avoid too many false alarms the detector threshold J_D has to be tuned accordingly. However, when tuning J_D there is a trade-off between the detection and false alarm rate one has to consider, while when there are attacks present one should also consider the impact of undetectable attacks when tuning J_D [9]. Typically, J_D is chosen such that rarely any false alarm happens. This means that the operator will not be suspicious if there are no alarms happening for a longer period of time. However, the trajectory of $S(k)$ is displayed in the control center, such that an operator could recognize suspicious behaviour by examining the displayed trajectory. This can also lead to the detection of the attacker and is represented as the human observer in Figure 1.

Assumption 2: The following conditions hold for the detector

1) $d\big(x_D(k), r(k)\big)$ is continuous in $x_D(k)$ and $r(k)$,
2) $S(k+1) = d\big(x_D(k), 0\big) \begin{cases} < S(k) & \text{if } x_D(k) \neq 0 \\ = 0 & \text{if } x_D(k) = 0 \end{cases}$,
3) $x_D(k) \to 0$ for $k \to \infty$, if $r(k) = 0 \quad \forall k$,
4) Set $x_D(k) = 0$, if $S(k) > J_D$.

The second and third condition are needed to guarantee that if we have perfect predictions of the received measurements,

i.e. $r(k)=0$, the detector state and output will approach zero without causing a false alarm. The fourth condition means that the detector is reset to its initial state, when an alarm has been triggered. Recall that a small detector output shows the operator that the system is behaving nominally.

Detectors that follow (6) and fulfil Assumption 2 are for example the χ^2, the MEWMA, the generalized MEWMA if parametrized appropriately, and the CUSUM detector.

Furthermore, under nominal conditions if $r(k)$ is a random variable then $S(k+1)$ is a random variable with probability density function $q_{k+1}(S)$ and support $\text{supp}\big(q_{k+1}(S)\big) \subseteq [0,\infty)$, where $\text{supp}\big(q(S)\big) := \{S \in \mathbb{R} : \ q(S) > 0\}$. Since $x_D(k)$ is an internal value of the detector, which is not transmitted over a network, it has been argued that $x_D(k)$ is confidential and only the operator has access to it (see for example [10]).

C. Attack Model

An attacker has penetrated the sensor network and can inject an additive signal $u_a(k) \in \mathbb{R}^{n_y}$ to the measurements, such that $\tilde{y}(k) = y(k) + u_a(k)$.

Assumption 3: The attacker has full system knowledge, i.e. knowledge of A, B, C, K, L, and Σ_r, and has access to the measurements $y(k)$, i.e. the attacker can use $y(k)$ in its design of $u_a(k)$. Furthermore, the attacker knows the detector equations $\theta(\cdot,\cdot)$ and $d(\cdot,\cdot)$ as well as the detector threshold J_D, but has neither access to $x_D(k)$, nor $S(k)$.

We make this assumption because we do not know about the attacker's system knowledge and capabilities. Therefore, we consider this a worst-case scenario. It is also reasonable to assume that the attacker has no access to $x_D(k)$ and $S(k)$, since these are internal variables not transmitted over the network. Further, under this assumption it has been shown (see for example [10]) that an attacker can design $u_a(k)$, such that $r(k) = a(k)$, where $a(k) \in \mathbb{R}^{n_y}$ is a vector chosen by the attacker. For the attack to remain undetected, i.e. $S(k+1) \leq J_D$ during the attack, $a(k)$ cannot be arbitrary but has to be designed appropriately. The lack of knowledge of $x_D(k)$ might decrease the attacker's attack space while remaining undetected. Therefore, previous research has either assumed that the attacker knows $x_D(k)$, when the attack happens [10] or assumed that the attack is designed such that it is remains undetected independently of $x_D(k)$ using a conservative bound [11]. For that reason, we want to investigate if an attacker can reduce its uncertainty about $x_D(k)$ without triggering an alarm or raising the operator's suspicion when $r(k) = a(k)$. Since the attacker has no access to $S(k+1)$, this represents an open-loop problem. In addition to that, another goal of the attacker is to maximize the average value of the estimation error $e(k) = x(k) - \hat{x}(k)$ in the Kalman filter. More specifically, the attacker concentrates on maximizing the maximum mean of the estimation error in the Kalman filter that belongs to safety critical states in the system (see Section III-B).

From the perspective of the defender it is important to know if the attacker is able to get to know $x_D(k)$, and if yes, what can one do to prevent that or at least prolong the process of obtaining the exact $x_D(k)$ and simultaneously mitigate the maximization of the average estimation error.

III. Attack Characterization

In this paper, we look at the case where $x_D(k)$ evolves with linear dynamics, i.e.

$$\begin{aligned} x_D(k+1) &= A_D x_D(k) + B_D r(k), \\ S(k+1) &= f\big(A_D x_D(k) + B_D r(k)\big) = d\big(x_D(k), r(k)\big). \end{aligned} \tag{7}$$

Here, $A_D \in \mathbb{R}^{n_D \times n_D}$ is Schur, $B_D \in \mathbb{R}^{n_D \times n_y}$ has full rank, and $n_D \leq n_y$. Further, $f(\cdot)$ is a vector norm on $\mathbb{R}^p$, A_D needs to be Schur, and $g(A_D) < 1$, where $g(\cdot)$ the matrix norm on $\mathbb{R}^{n_D \times n_D}$ induced by $f(\cdot)$, such that the first three detector conditions in Assumption 2 are fulfilled. Without loss of generality we assume the attack starts at $k=0$, such that $r(k) = a(k)\ \forall k \geq 0$ and $x_D(0)$ is unknown to the attacker. Since the dynamics are linear we can without loss of generality rewrite the detector state as $x_D(k) = x_{D,r}(k) + x_{D,a}(k)$, where

$$\begin{aligned} x_{D,a}(k+1) &= A_D x_{D,a}(k) + B_D a(k) \\ x_{D,r}(k+1) &= A_D x_{D,r}(k), \end{aligned}$$

with $x_{D,r}(0) = x_D(0)$ and $x_{D,a}(0) = 0$. Here, $x_{D,a}(k)$ is governed by the attack signal, while $x_{D,r}(k)$ is an autonomous system, which is governed by the initial state of the detector. Since A_D is Schur, $x_{D,r}(k) \to 0$ as $k \to \infty$. This means that $x_{D,a}(k)$ can be seen as the estimate of $x_D(k)$ at time step k and $x_{D,a}(k) \to x_D(k)$ as $k \to \infty$.

To have a good estimate, i.e. reduce the uncertainty, at time step N, we want

$$\begin{aligned} ||x_D(N) - x_{D,a}(N)||_2 = ||x_{D,r}(N)||_2 &\leq \gamma \\ \Leftrightarrow \big|\big|A_D^N x_D(0)\big|\big|_2 &\leq \gamma \end{aligned}$$

where $\gamma > 0$ is close to zero and $||o||_2$ represents the Euclidean norm of o. Since $x_D(0)$ is unknown, we obtain an upper bound $S_{\text{up}} = \max_x ||x||_2$ subject to $x \in \{y \in \mathbb{R}^{n_y} : f(y) \leq J_D\}$. With that we choose N, such that

$$\big|\big|A_D^N\big|\big|_2 = \sigma_{\max}\big(A_D^N\big) \leq \frac{\gamma}{S_{\text{up}}}, \tag{8}$$

where $\sigma_{\max}(C)$ is the maximum singular value of matrix C.

Remark 1: We see that the slower $\sigma_{\max}(A_D^k)$ approaches zero as $k \to \infty$ the more time it takes for the attacker to obtain a close estimate of the $x_D(k)$. Hence, a defender can consider this fact, when designing the detector.

The attacker not only wants to reduce its uncertainty about $x_D(k)$, but also wants to remain undetected by the detector. Therefore, we look now at the condition for the attacker to remain undetected. Since $f(\cdot)$ is a vector norm, we can determine the following condition to avoid detection.

$$\begin{aligned} S(k) &= f\big(x_{D,r}(k) + x_{D,a}(k)\big) \\ &\leq f\big(x_{D,r}(k)\big) + f\big(x_{D,a}(k)\big) \\ &\leq g\big(A_D^k\big) f\big(x_{D,r}(0)\big) + f\big(x_{D,a}(k)\big) \\ &\leq g\big(A_D^k\big) J_D + f\big(x_{D,a}(k)\big) \leq J_D, \\ \Rightarrow S_a(k) &= f\big(x_{D,a}(k)\big) \leq J(k), \end{aligned}$$

where $J(k) = J_D - g\big(A_D^k\big) J_D > 0$ for all $k > 0$. We see that if $S_a(k) \leq J(k)$, then the attack remains undetected. Note that $J(k) \to J_D$ as $k \to \infty$. We can interpret $x_{D,a}(k)$ and $S_a(k)$ as a

virtual detector with threshold $J(k)$ that the attacker initializes at $x_{D,a}(0)=0$ and uses to design its undetectable attack.

Let us summarize these results in a proposition and then discuss how to design $\{a(k)\}_{k=0}^{N-1}$.

Proposition 1: An attacker can reconstruct the detector state with accuracy γ, in N time steps, where N is such that $\sigma_{\max}\left(A_D^N\right) \leq \frac{\gamma}{S_{\text{up}}}$ with $S_{\text{up}} = \max_{x\in\{y\in\mathbb{R}^{n_y}:f(y)\leq J_D\}} ||x||_2$ is fulfilled. The attacker can simultaneously inject attacks $a(k)$ satisfying $S_a(k+1) = d\left(x_{D,a}(k), a(k)\right) \leq J(k+1)$ without triggering alarms.

A simple way for the attacker to choose the residual is $a(k)=0$ for $k\in\{0,\cdots,N-1\}$, since then $x_{D,a}(k)=0$ for all $k\geq 0$, which implies that $S_a(k)=0\leq J(k)$ for all $k\geq 0$. However, this leads to suspicious behaviour in $S(k+1)$, for example an exponential decay of $S(k+1)$, which might raise an operator's suspicion when seeing this on the display in the control center. Another way is to not change the measurements and just observe $r(k)$ and feed it into $x_{D,a}(k)$. The advantage is that $S(k+1)$ behaves exactly as in the nominal case, but without knowledge of $x_D(k)$ any $r(k)$ might lead to an alarm, which is considered as a false-alarm under nominal conditions. A third option is to make the alarm look like a false alarm, by inducing a spike in one element of $r(k)$, such that $x_D(k)$ is reset to zero. However, since the attacker is present in the system, this "false alarm" in the last two strategies might lead to the detection of the attacker.

Therefore, an attacker needs to design $a(k)$ in such a way that under attack $S(k+1)$ approximately has probability density $q_{k+1}(S)$, but no alarms are caused. Since $x_{D,a}(k)\to x_D(k)$ as $k\to\infty$, we also get $S_a(k)\to S(k)$ as $k\to\infty$. Therefore, we look at the virtual detector $S_a(k)$ instead of $S(k)$, since the attacker has no direct access to $S(k)$. This means when $S(k)$ has a probability density function $q_k(S)$, which changes for a given $x_D(k-1)$ then we assume that $S_a(k)$ has the same probability density function $q_k(S)$ but given $x_{D,a}(k-1)$. This can be formulated as two problems that need to be solved at each time step k.

Problem 1: Find a probability density function $p_{k+1}(S)$, such that $\text{supp}\left(p_{k+1}(S)\right) = [0, J(k+1)]$ (no alarms) and $p_{k+1}(S)$ resembles $q_{k+1}(S)$ as closely as possible.

Problem 2: Draw a sample s_{k+1} from the probability distribution with probability density function $p_{k+1}(S)$ and design $a(k)$ such that $S_a(k+1)=s_{k+1}$ and the maximum average estimation error of the critical states is maximized.

In the following, we propose solutions to these two problems.

A. How to characterize $p_k(S)$ (Problem 1)

Kullback *et al.* [14] defined the average information gain of each observation to distinguish between a hypothesis with density function $p(S)$ and a hypothesis with density function $q(S)$ as $D_{KL}(p||q)=\int p(S)\ln\left(\frac{p(S)}{q(S)}\right)dS$, which is known as the Kullback-Leibler (KL) divergence. Furthermore, $D_{KL}(p||q)$ is convex in the pair of its arguments. Therefore, it comes quite natural that we try to minimize the average information gain $D_{KL}(p_k||q_k)$ to find $p_k(S)$,

$$\min_{p_k(S)} \int_0^{J(k)} p_k(S)\ln\left(\frac{p_k(S)}{q_k(S)}\right)dS$$
$$\text{s.t.} \begin{cases} p_k(S)\geq 0 \quad \forall S\in[0,J(k)] \\ p_k(S)=0 \quad \forall S\notin[0,J(k)] \\ \int_0^{J(k)} p_k(S)dS = 1 \\ \text{more convex constraints on } p_k(S) \end{cases}. \tag{9}$$

The first three constraints are necessary such that $p_k(S)$ is a probability density function. One can also impose more convex constraints on $p_k(S)$ which preserve the convexity of the problem. For example, we can impose a constraint on the mean $\int_0^{J(k)} Sp_k(S)dS$ or the second raw moment $\int_0^{J(k)} S^2 p_k(S)dS$ as well.

In this paper we only look at the case where no additional constraints are imposed. Then, we need to solve

$$\min_{p_k(S)} \int_0^{J(k)} p_k(S)\ln\left(\frac{p_k(S)}{q_k(S)}\right)dS$$
$$\text{s.t.} \begin{cases} p_k(S)\geq 0 \quad \forall S\in[0,J(k)] \\ p_k(S)=0 \quad \forall S\notin[0,J(k)] \\ \int_0^{J(k)} p_k(S)dS = 1 \end{cases}. \tag{10}$$

Proposition 2: The optimizer to (10) is

$$p_k^*(S) = \begin{cases} \frac{q_k(S)}{\int_0^{J(k)} q_k(S)dS} & S\in[0,J(k)] \\ 0 & \text{otherwise} \end{cases}, \tag{11}$$

i.e. the truncated version of $q_k(S)$ is the optimal solution.

Proof: Let $\lambda\in\mathbb{R}$ be a Lagrange multiplier and the Lagrangian be

$$\begin{aligned} L(p,\lambda) &= \int_0^{J(k)} p_k(S)\ln\left(\frac{p_k(S)}{q_k(S)}\right)dS + \lambda\left(\int_0^{J(k)} p_k(S)dS - 1\right) \\ &= \int_0^{J(k)} p_k(S)\ln\left(\frac{p_k(S)}{q_k(S)}\right) + \lambda\left(p_k(S) - \frac{1}{J(k)}\right)dS \\ &= \int_0^{J(k)} l\left(p_k(S),\lambda\right)dS \end{aligned}$$

A necessary condition for optimality (see [15]) is

$$\frac{d}{dp_k(S)} l\left(p_k(S),\lambda\right)\Big|_{p_k(S)=p_k^*(S)} = 0.$$

Solving for $p_k^*(S)$ leads to

$$p_k^*(S) = \begin{cases} e^{-1-\lambda}q_k(S) & \forall S\in[0,J(k)] \\ 0 & \forall S\notin[0,J(k)] \end{cases},$$

where we already incorporated the first two constraints of (10). Now we use the last constraint to find

$$\lambda = -1 + \ln\left(\int_0^{J(k)} q_k(S)dS\right),$$

which results in $p_k^*(S)$. ■

B. How to characterize $a(k)$ (Problem 2)

Once we determined $p_{k+1}(S)$, we take a sample from this distribution. Let the obtained sample be s_{k+1}. Now we want to design $a(k)$ such that $S_a(k+1) = f\big(A_D x_{D,a}(k) + B_D a(k)\big) = s_{k+1}$. As mentioned before the attacker also wants to maximize the average estimation error $e(k) = x(k) - \hat{x}(k)$ of the operator. The dynamics of the estimation error are given by

$$e(k+1) = Ae(k) - L\bar{r}(k) + w(k), \tag{12}$$

where $\bar{r}(k) = \Sigma_r^{\frac{1}{2}} r(k) = \Sigma_r^{\frac{1}{2}} a(k)$ under the attack. Without loss of generality, we can write $e(k) = e_n(k) + e_a(k)$, where $e_n(k)$ represents the part of the error that is influenced by the process noise $w(k)$ and $e_a(k)$ is the part of the estimation error that is driven by the attack signal. Further, $e_n(0) = e(0)$ and $e_a(0) = 0$. We can interpret $e_a(k)$ as the average value of the estimation error at time step k. Let us introduce the estimation error of critical states as $e_{crit}(k) = T_c e(k)$, where $T_c \in \mathbb{R}^{n_c \times n_x}$ is a matrix that extracts the critical estimation errors of $e(k)$ and $n_c \leq n_x$. This could for example be the estimation error of the pressure in a closed container, which might explode if the pressure is too large. Therefore, the attacker wants to maximize the maximum estimation error of these critical states. The optimization problem to find $a(k)$ becomes then

$$\begin{aligned} \mathscr{I}_e = \max_{a(k)} ||T_c e_a(k+1)||_\infty = \max_{a(k)} \big|\big|T_c A e_a(k) - T_c L \Sigma_r^{\frac{1}{2}} a(k)\big|\big|_\infty \\ \text{s.t. } S_a(k+1) = f\big(A_D x_{D,a}(k) + B_D a(k)\big) = s_{k+1}, \end{aligned} \tag{13}$$

where both $e_a(k)$, $x_{D,a}(k)$, and s_{k+1} are known to the attacker.

Before we introduce the solution to (13), we define the dual norm of a vector norm [16].

Definition 1: The dual norm of a vector norm $f(x)$ in $\mathbb{R}^n$ is defined as

$$f^{\mathscr{D}}(z) := \max_x \left|z^T x\right| \text{ s.t. } f(x) = 1,$$

where $z \in \mathbb{R}^n$.

Now we introduce an intermediate result for solving (13).

Lemma 1: The optimal value $\mathscr{I}$ of

$$\max_{\bar{a}} |\bar{c}^T \bar{a} + \bar{d}| \text{ s.t. } f(\bar{a}) = s, \tag{14}$$

where $s \geq 0$, $\bar{d} \in \mathbb{R}$, $\bar{a}, \bar{c} \in \mathbb{R}^{n_D}$, is given by

$$\mathscr{I} = \max\left(|f^{\mathscr{D}}(\bar{c})s + \bar{d}|, |-f^{\mathscr{D}}(\bar{c})s + \bar{d}|\right) \tag{15}$$

with the maximizer

$$\bar{a}^* = \arg\max_{\bar{a}} (-1)^j c^T \bar{a} \text{ s.t. } f(\bar{a}) \leq s. \tag{16}$$

Here, $j = 2$ if $\left|f^{\mathscr{D}}(\bar{c})s + \bar{d}\right| \geq \left|-f^{\mathscr{D}}(\bar{c})s + \bar{d}\right|$ and $j = 1$ otherwise.

Proof: We first split (14) into two optimization problems, one that maximizes and one that minimizes $\bar{c}^T \bar{a} + \bar{d}$ under the given constraint respectively. The larger absolute value of the optimal values of these two problems gives us the solution to (14). Note that $\bar{d}$ is a scalar and therefore the optimizer of these two problems, will maximize or minimize $\bar{c}^T \bar{a}$, respectively. Definition 1 gives us that $\max_{f(\bar{a})=s} \left|\bar{c}^T \bar{a}\right| = f^{\mathscr{D}}(\bar{c})s$, from which (15) readily follows. Since the optimizer lies on the boundary of the constraint set, we replace the equality constraint of (14) with an inequality constraint to obtain the convex optimization (16). ■

Theorem 1: The solution $\mathscr{I}_e$ of (13) is given by

$$\mathscr{I}_e = \max_{i \in \{1,\cdots,n_c\}} \max\left(\left|f^{\mathscr{D}}(\bar{c}_i)s_{k+1} + \bar{d}_i\right|, \left|-f^{\mathscr{D}}(\bar{c}_i)s_{k+1} + \bar{d}_i\right|\right),$$

and the corresponding attack vector can be found as

$$a(k) = B_D^{\dagger}\big(\bar{a}^* - A_D x_{D,a}(k)\big) \tag{17}$$

with $\bar{a}^*$ being the optimizer of the convex problem

$$\bar{a}^* = \arg\max_{\bar{a}} (-1)^{j_{i^*}} \bar{c}_{i^*}^T \bar{a} \text{ s.t. } f(\bar{a}) \leq s_{k+1}$$

where $\bar{a} \in \mathbb{R}^{n_D}$, $i^* \in \{1,\cdots,n_c\}$ denotes an element of $T_c e_a(k+1)$ for which $\mathscr{I}_e$ is achieved, and $j_{i^*} = 2$ if $\left|f^{\mathscr{D}}(\bar{c}_{i^*})s_{k+1} + \bar{d}_{i^*}\right| \geq \left|-f^{\mathscr{D}}(\bar{c}_{i^*})s_{k+1} + \bar{d}_{i^*}\right|$ and $j_{i^*} = 1$ otherwise.

Here, $\bar{c}_i^T = -t_i^T L \Sigma_r^{\frac{1}{2}} B_D^{\dagger}$, $\bar{d}_i = t_i^T \left(Ae_a(k) + L\Sigma_r^{\frac{1}{2}} B_D^{\dagger} A_D x_{D,a}(k)\right)$, and t_i^T is the ith row of T_c. Further, $B_D^{\dagger}$ denotes the Moore-Penrose pseudoinverse of B_D such that $B_D B_D^{\dagger} = I_{n_D}$ because B_D is full rank and $n_D \leq n_y$. Here, I_o represents the o dimensional identity matrix.

Proof: We exploit that $||T_c e_a(k+1)||_\infty = \max_{i \in \{1,\cdots,n_c\}} \left|t_i^T e_a(k+1)\right|$, where $t_i^T e_a(k+1)$ represents the estimation error of the ith critical state. This approach has also been used in [4]. Therefore, we can solve n_c problems of the form

$$\begin{aligned} \max_{a(k)} \left|t_i^T A e_a(k) - t_i^T L \Sigma_r^{\frac{1}{2}} a(k)\right| \\ \text{s.t. } f\big(A_D x_{D,a}(k) + B_D a(k)\big) = s_{k+1}, \end{aligned} \tag{18}$$

where $i \in \{1,\cdots,n_c\}$ and pick $a(k)$ which results in the maximal objective value of all of these problems. Introducing $\bar{a} = A_D x_{D,a}(k) + B_D a(k)$, we reformulate (18) as

$$\max_{\bar{a}} \left|\bar{c}_i^T \bar{a} + \bar{d}_i\right| \text{ s.t. } f(\bar{a}) = s_{k+1}, \tag{19}$$

which represents n_c problems of the form presented in Lemma 1. Therefore, we can use Lemma 1 to determine both $\mathscr{I}_e$, $\bar{a}$ and with that $a(k)$. ■

Remark 2: If $f(x) = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$, where $1 \leq p \leq \infty$, and x_i is the ith element of x, then $f^{\mathscr{D}}(x) = \left(\sum_i |x_i|^q\right)^{\frac{1}{q}}$ such that $\frac{1}{p} + \frac{1}{q} = 1$. This is a result of the Hölder inequality (see [16]).

Remark 3: One can also think of solutions, which take other objectives into account when designing $a(k)$ at each time step. However, we chose this objective, because it maximizes the estimation error of the critical state in the sense of the maximum norm and we are able to find an analytical solution.

IV. Application to The MEWMA Detector

Now we apply the previously presented procedure to the MEWMA detector and give a numerical example. Here, we assume that no extra constraints on $p_k(S)$ are imposed.

A. The MEWMA detector

The MEWMA detector in [8] is given by

$$\begin{aligned} x_D(k+1) &= \beta r(k) + (1-\beta)x_D(k) \\ \tilde{S}(k+1) &= \frac{2-\beta}{\beta}||x_D(k+1)||_2^2, \end{aligned} \tag{20}$$

where $\beta \in (0,1]$. If $\tilde{S}(k+1) \leq \tilde{J}_D$ no alarm is triggered, where $\tilde{J}_D \in \mathbb{R}_{\geq 0}$ and if an alarm happens the detector state is reset to zero. The MEWMA detector as defined in [8], does not fit the detector model in (7), but we can rewrite it as

$$\begin{aligned} x_D(k+1) &= \beta r(k) + (1-\beta)x_D(k) \\ S(k+1) &= ||x_D(k+1)||_2 \end{aligned} \tag{21}$$

and use $J_D = \sqrt{\frac{\beta}{2-\beta}\tilde{J}_D}$ as the new detector threshold. This now fits (7) with $A_D = (1-\beta)$, $B_D = \beta$, $f(\cdot)$ being the Euclidean norm, and $g(\cdot) = \sigma_{\max}(\cdot)$.

Recall, $x_D(0)$ is unknown to the attacker. Since the dynamics are linear, we split the MEWMA detector into two parts, $x_{D,a}(k)$ and $x_{D,r}(k)$, so that $x_D(k) = x_{D,a}(k) + x_{D,r}(k)$, where

$$\begin{aligned} x_{D,a}(k+1) &= \beta a(k) + (1-\beta)x_{D,a}(k) \\ x_{D,r}(k+1) &= (1-\beta)x_{D,r}(k), \end{aligned}$$

$k \geq 0$, $x_{D,r}(0) = x_D(0)$, and $x_{D,a}(0) = 0$.

Now we determine the attack duration N according to (8).

Proposition 3: The uncertainty of the MEWMA detector's state at time step N is smaller than $\gamma > 0$, i.e. $||x_D(N) - x_{D,a}(N)||_2 \leq \gamma$ if

$$N \geq \left\lceil \frac{\ln(\frac{\gamma}{J_D})}{\ln(1-\beta)} \right\rceil, \tag{22}$$

where $\lceil x \rceil$ rounds x up to the next larger integer value. Note that $\gamma \leq J_D$ for $N \geq 0$.

Proof: Since $A_D = 1-\beta$, we see that $\sigma_{\max}(A_D^N) = (1-\beta)^N$. Further, we determine that $S_{up} = J_D$. With that we solve (8) for N and obtain inequality (22). ■

The attacker can launch an attack for N time steps such that the initial detector state $x_D(0)$ decreased sufficiently so that the attacker's uncertainty about $x_D(k)$ at time step N is small, i.e. $x_D(N) \approx x_{D,a}(N)$. Further, for the attack to remain undetected we obtain $J(k) = J_D(1-(1-\beta)^k)$, because $g(A_D^k) = (1-\beta)^k$.

Now that we have determined N and $J(k)$ let us determine the probability density function $p_k(S)$ by finding $q_k(S)$ under nominal conditions. Here, we change the procedure of Section III slightly and look at

$$\frac{1}{\beta^2}S(k+1)^2 = ||r(k) + \frac{1-\beta}{\beta}x_D(k)||_2^2$$

instead of $S(k+1)$ because in the nominal case this follows a noncentral χ^2 distribution with n_y degrees of freedom and noncentrality parameter $\lambda(k+1) = \left(\frac{1-\beta}{\beta}\right)^2 x_D(k)^T x_D(k)$ at each time step.

Therefore according to Proposition 2 we design $p_{k+1}(S)$ as a truncated noncentral χ^2 distribution with n_y degrees of freedom, noncentrality parameter $\lambda_a(k+1) = \left(\frac{1-\beta}{\beta}\right)^2 x_{D,a}(k)^T x_{D,a}(k)$ and support $\text{supp}(p_{k+1}(S)) = \left[0, \frac{1}{\beta^2}J(k+1)^2\right]$.

After we draw a sample s_{k+1} from the truncated noncentral χ^2 distribution $p_{k+1}(S)$, we use (13) to determine $a(k)$, which for the MEWMA case looks as follows

$$\begin{aligned} \mathscr{I}_e = &\max_{a(k)} ||T_c e_a(k+1)||_\infty \\ &\text{s.t. } ||\beta a(k) + (1-\beta)x_{D,a}(k)||_2 = \beta\sqrt{s_{k+1}}. \end{aligned} \tag{23}$$

Corollary 1: The impact for the MEWMA is

$$\mathscr{I}_e^M = \max_{i \in \{1,\cdots,n_c\}} \max\left(\left| ||\bar{c}_i||_2 \beta\sqrt{s_{k+1}} + \bar{d}_i \right|, \left| -||\bar{c}_i||_2 \beta\sqrt{s_{k+1}} + \bar{d}_i \right| \right)$$

for the attack vector

$$a(k) = (-1)^{j_{i^*}} \frac{\bar{c}_{i^*}}{||c_{i^*}||_2}\sqrt{s_{k+1}} - \frac{1-\beta}{\beta}x_{D,a}(k),$$

where i^* is an index that results in $\mathscr{I}_e^M$, $\bar{c}_i^T = -\frac{1}{\beta}t_i^T L \Sigma_r^{\frac{1}{2}}$, $\bar{d}_i = t_i^T\left(Ae_a(k) + \frac{1-\beta}{\beta}L\Sigma_r^{\frac{1}{2}}x_{D,a}(k)\right)$, and $j_{i^*} = 2$ if $\left| ||\bar{c}_i||_2\beta\sqrt{s_{k+1}} + \bar{d}_i \right| \geq \left| -||\bar{c}_i||_2\beta\sqrt{s_{k+1}} + \bar{d}_i \right|$ and $j_{i^*} = 1$ otherwise.

This follows readily from Theorem 1.

B. Numerical Example

Let us now look at a numerical example to verify the procedure for the MEWMA detector.

For the simulation we use the linearized twelve dimensional reduced order model of a 76 story building given in [17]. We use the model with $n_y = 20$ measurements and further discretize it with a sampling period of $T_s = 0.01\,\text{s}$ to design the LQG controller. For the MEWMA detector, we use $\beta = 0.2$ and $\tilde{J}_D = 40$, which leads to an average time between false alarms of approximately 284 time steps. To reduce the uncertainty about $x_D(k)$, we choose $\gamma = 10^{-6}$ such that we get an attack length of $N = 66$ time steps according to (22). We let the system run for 100 time steps initially and then start the attack at $k = 100$, to obtain a comparison of $\tilde{S}(k)$ before and after the attack. Further, for this simulation we use $T_c = I_{n_x}$.

The upper plot in Figure 2 shows one simulation run of the trajectory of $\tilde{S}(k)$ for the MEWMA detector as described in (20) before and after the attack and the trajectory of $\tilde{S}_a(k)$ that is the virtual MEWMA detector the attacker uses to design its undetectable attack. We see that after the attack starts the trajectory $\tilde{S}(k)$ still is random and does not show any obvious irregularities to the bare human eye. Furthermore, the attack is not detected since the alarm threshold $\tilde{J}_D = 40$ is never crossed and we observe that $\tilde{S}_a(k) \to \tilde{S}(k)$ as time progresses. This shows us that the attacker's estimate of $x_D(k)$ becomes more accurate over time. Finally, we look at the accuracy of the estimate at the end of the attack. We have $||x_D(166) - x_{D,a}(166)||_2 = 6.2572 \cdot 10^{-7}$. As desired, the uncertainty is smaller than $\gamma = 10^{-6}$.

The lower plot of Figure 2 shows the average trajectory of $||T_c e_a(k)||_\infty$ over 10000 simulations. Here, we see that the maximum estimation error is on average increasing over time.

Therefore, we verified that an attacker with access to and control over the measurements is able to break the confidentiality of the internal state of the MEWMA detector and to simultaneously increase the maximum estimation error of the critical states.

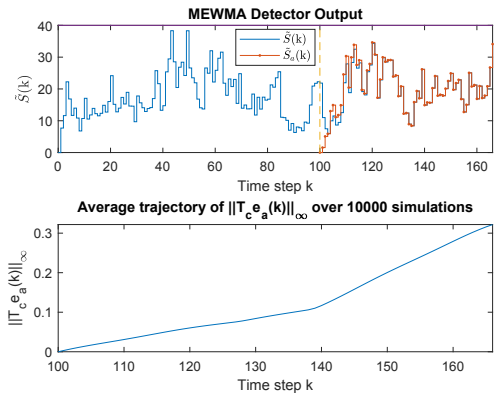

Fig. 2. The upper plot shows how $\tilde{S}(k)$ behaves before and after the attack starting at $k = 100$. The lower plot shows the average trajectory of $||T_c e_a(k)||_\infty$ over 1000 simulations

V. CONCLUSIONS

In this paper, we investigated ways how an attacker that has access to all measurements and can change them arbitrarily is able to reduce its uncertainty about the detector's internal state that evolves with linear dynamics. The proposed procedure takes the statistics of the detector output signal into account such that an operator will not notice irregularities if it looks at the detector output even if no alarm is triggered. Further, the attacker is able to maximize the maximum norm of the average estimation error of the critical states in the Kalman filter at each time step.

There are several directions of future work we have with this work. In our work, we showed how an attacker can break the confidentiality of the detector's internal state. Therefore it is important to look into defence mechanisms such that the estimation of the detector's state becomes more difficult or even impossible for the attacker. Further, here we only looked at detector which have linear dynamics. Therefore, extending this procedure to more general anomaly detector models is of interest as well. This might result in novel anomaly detectors for which the confidentiality cannot be broken.

REFERENCES

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

[2] R. M. Ferrari and A. M. Teixeira, "Detection and isolation of replay attacks through sensor watermarking," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 7363 – 7368, 2017, 20th IFAC World Congress.

[3] P. Hespanhol, M. Porter, R. Vasudevan, and A. Aswani, "Statistical watermarking for networked control systems," in *2018 Annual American Control Conference (ACC)*, June 2018, pp. 5467–5472.

[4] J. Milosevic, D. Umsonst, H. Sandberg, and K. H. Johansson, "Quantifying the impact of cyber-attack strategies for control systems equipped with an anomaly detector," in *European Control Conference 2018*, June 2018.

[5] F. Pasqualetti, F. Dörfler, and F. Bullo, "Control-theoretic methods for cyberphysical security: Geometric principles for optimal cross-layer resilient control systems," *IEEE Control Systems*, vol. 35, no. 1, pp. 110–127, 2015.

[6] Z. Wang, F. Harirchi, D. Anand, C. Y. Tang, J. Moyne, and D. Tilbury, "Conflict-driven hybrid observer-based anomaly detection," in *2018 Annual American Control Conference (ACC)*, June 2018, pp. 5793–5800.

[7] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[8] C. A. Lowry, W. H. Woodall, C. W. Champ, and S. E. Rigdon, "A multivariate exponentially weighted moving average control chart," *Technometrics*, vol. 34, no. 1, pp. 46–53, 1992.

[9] D. I. Urbina, J. A. Giraldo, A. A. Cárdenas, N. O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, and H. Sandberg, "Limiting the impact of stealthy attacks on industrial control systems," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: ACM, 2016, pp. 1092–1105.

[10] C. Murguia and J. Ruths, "Cusum and chi-squared attack detection of compromised sensors," in *2016 IEEE Conference on Control Applications (CCA)*, Sept 2016, pp. 474–480.

[11] D. Umsonst and H. Sandberg, "Anomaly detector metrics for sensor data attacks in control systems," in *2018 Annual American Control Conference (ACC)*, June 2018, pp. 153–158.

[12] D. M. Hawkins, S. Choi, and S. Lee, "A general multivariate exponentially weighted moving-average control chart," *Journal of Quality Technology*, vol. 39, no. 2, pp. 118–125, 2007.

[13] I. Hwang, S. Kim, Y. Kim, and C. E. Seah, "A survey of fault detection, isolation, and reconfiguration methods," *IEEE Transactions on Control Systems Technology*, vol. 18, no. 3, pp. 636–653, May 2010.

[14] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[15] D. G. Luenberger, *Optimization by Vector Space Methods*, 1st ed. New York, NY, USA: John Wiley & Sons, Inc., 1997.

[16] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. New York, NY, USA: Cambridge University Press, 2012.

[17] J. N. Yang, A. K. Agrawal, B. Samali, and J.-C. Wu, "Benchmark problem for response control of wind-excited tall buildings," *Journal of Engineering Mechanics*, vol. 130, no. 4, pp. 437–446, 2004.