

# On the confidentiality of controller states under sensor attacks <sup>★</sup>

David Umsonst <sup>a</sup>, Henrik Sandberg <sup>a</sup>,

<sup>a</sup>*Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden*

---

## Abstract

With the emergence of cyber-attacks on control systems it has become clear that improving the security of control systems is an important task in today's society. We investigate how an attacker that has access to the measurements transmitted from the plant to the controller can perfectly estimate the internal state of the controller. This attack on sensitive information of the control loop is, on the one hand, a violation of the privacy, and, on the other hand, a violation of the security of the closed-loop system if the obtained estimate is used in a larger attack scheme. Current literature on sensor attacks often assumes that the attacker has already access to the controller's state. However, this is not always possible. We derive conditions for when the attacker is able to perfectly estimate the controller's state. These conditions show that if the controller has unstable poles a perfect estimate of the controller state is not possible. Moreover, we propose a defence mechanism to render the attack infeasible. This defence is based on adding uncertainty to the controller dynamics. We also discuss why an unstable controller is only a good defence for certain plants. Finally, simulations with a three-tank system verify our results.

*Key words:* Cyber-physical security; Privacy; Linear control systems; Kalman filters; Algebraic Riccati equations; Discrete-time systems.

---

## 1 Introduction

The smart grid and intelligent transportation systems are two prime examples of cyber-physical systems, where physical processes are controlled over communication networks and with digital computers. The interconnection of the physical and cyber domain promises great advantages in the performance and capabilities of cyber-physical systems. However, with the introduction of communication networks and computational devices, the controlled processes become vulnerable to cyber-attacks. Documented cyber-attacks such as the Stuxnet attack on an Iranian uranium enrichment facility (Kushner, 2013), the cyber attack on a German steel mill (Lee et al., 2014), and the BlackEnergy attack on the Ukrainian power grid (Lee et al., 2016) show that these attacks are not a futuristic concept but already happening.

Teixeira et al. (2015) define a cyber-physical attack space that is spanned by the attacker's disclosure and disruptive resources as well as its model knowledge. Disclosure resources enable the attacker to gather information about the system and, therefore, break its confidentiality. These *disclosure attacks* can, for example, be used to increase the attacker's model knowledge. Disruptive resources, on the other hand, let the attacker launch both deception and denial of service attacks, which affect the integrity and the availability of measurement and actuator signals, respectively. Several attacks can be mapped into this attack space, for example replay attacks, where well-behaved sensor measurements are replayed, while the actuator signals are changed.

Although many attack strategies have been investigated, the analysis of sensor attacks has gained popularity in the last decade. A goal of the sensor attacks is to remain undetected by the anomaly detector of the operator, while changing the measurements. Figure 1 shows the block diagram of the cyber-physical system under a sensor attack. Here, the dashed line going to the attacker corresponds to the disclosure resources of the attacker. The disclosure resources can be used to gather more information about the closed-loop system,

---

<sup>★</sup> This work is supported in part by the Swedish Research Council (grant 2016-00861), the Swedish Energy Agency (project LarGo!), and the Swedish Civil Contingencies Agency (project CERES).

*Email addresses:* [umsonst@kth.se](mailto:umsonst@kth.se) (David Umsonst), [hsan@kth.se](mailto:hsan@kth.se) (Henrik Sandberg).

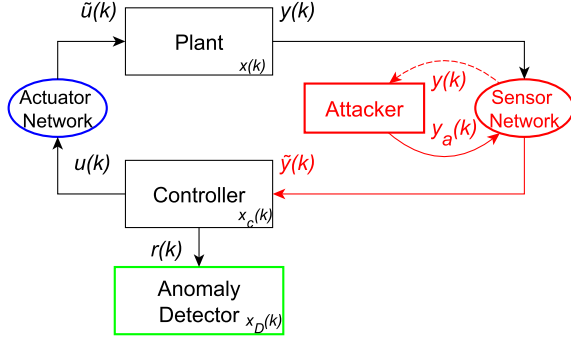


Fig. 1. Block diagram of the closed-loop system equipped with an anomaly detector under a sensor attack

for example, about the internal states of the plant  $x(k)$ , the controller  $x_c(k)$ , or the anomaly detector  $x_D(k)$ . The disruptive resources are denoted by  $y_a(k)$  and are used to change the values of the measurements from  $y(k)$  to  $\tilde{y}(k) = y(k) + y_a(k)$ . Mo and Sinopoli (2010) look into integrity attacks on sensors and define a notion of perfectly attackable systems, while Cárdenas et al. (2011) analyse two different detectors and three sensor attack strategies. Another approach is to maximize the error covariance matrix of a state estimator with a sensor attack as it is done in Guo et al. (2018). In Murguía and Ruths (2016), a sensor attack strategy is proposed which replaces the residual signal  $r(k)$ , which is the input to the anomaly detector (Fig. 1), with a signal designed by the attacker. This attack strategy is then used to look at the impact under the two anomaly detectors investigated in Cárdenas et al. (2011). In Umsonst and Sandberg (2019), it was shown how an attacker using this attack strategy is able to break the confidentiality of the internal anomaly detector state  $x_D(k)$ .

What connects all these papers on sensor attacks is that the attacker needs to have *exact* knowledge about the internal state  $x_c(k)$  of the controller, when the attack starts. Mo and Sinopoli (2010) assume that the initial system state is zero in order to determine the undetectable attack, while the other papers assume that the controller state is known to the attacker. Therefore, this paper investigates a missing piece that is often taken for granted in sensor attacks, namely the broken confidentiality of the controller's internal state. More precisely, we examine if an attacker with full model knowledge listening to the sensor measurements is able to break the confidentiality of the controller. This confidentiality attack can have two purposes. One purpose might be that the attacker is curious and wants to follow the activity of the control centre. The other purpose might be that it is one step in a more complex attack scheme. This step represents gathering information of the plant and its controller, which is then used in later steps to attack the system. We can interpret this as a first step the at-

tacker needs to perform to execute the attacks proposed by the papers mentioned above.

### 1.1 Contributions

The contribution of our work is three-fold. Firstly, we provide a rigorous analysis of whether or not an attacker with full-model knowledge and access to all sensors is able to *perfectly* estimate the internal state of the output-feedback controller. Although it may seem obvious that such a powerful attacker is able to estimate the state perfectly, we show that if the operator uses an unstable controller the attacker is not able to do so. We further classify all gains for a linear time-invariant observer the attacker could use to achieve a perfect estimate in case of a stable output-feedback controller. The second contribution provides a defence mechanism against this disclosure attack. This mechanism proposes to add some uncertainty to the controller's input, which can be interpreted as a watermarking scheme. Furthermore, we discuss when an unstable controller is an appropriate defence mechanism. The third contribution is the verification of our theoretical results with simulations of a three-tank system under attack.

### 1.2 Related work

Most of the research on the security of cyber-physical systems has focused on the integrity and availability of data, according to Lun et al. (2019). For example, all the previously mentioned papers on sensor attacks except Umsonst and Sandberg (2019) consider integrity attacks.

Other work on the confidentiality of control systems can be found in, for example, Xue et al. (2014), Yuan and Mo (2015), and Dibaji et al. (2018). What distinguishes this paper from these results is that we focus on a general linear system structure, while Xue et al. (2014) investigate the confidentiality of a special structured linear system. Further, we focus on the confidentiality of the controller's internal state. However, Xue et al. (2014) consider the confidentiality of the whole system state, while Dibaji et al. (2018) look into the confidentiality of controller gains and in Yuan and Mo (2015) the attacker wants to identify the controller structure. In Yuan and Mo (2015), it is shown that an appropriate controller design can lead to confidentiality. We also show that a certain type of controller, in our case an unstable controller, leads to confidentiality regarding sensitive information of the controller. It is interesting that an appropriate controller design can preserve the controller's confidentiality.

Another recent research direction is to use homomorphic encryption to ensure the security and privacy of control systems (Kogiso and Fujita, 2015; Farokhi et al.,

2017). Based on encrypted sensor measurements the controller determines an encrypted control signal, which is decrypted at the actuator, i.e., the feedback loop operates on encrypted signals. The use of encrypted signals guarantees that, even if the attacker estimates the controller state, the estimate is not useful to the attacker, due to the encryption. In our approach, we use an artificial uncertainty instead of encryption techniques to preserve the confidentiality of the controller.

Defending the cyber-physical system against attacks by introducing an artificial uncertainty is also done in the work using watermarking. Watermarking of the actuator signal has been considered as a defence mechanism, for example, against replay attacks (Mo et al., 2015) or sensors attacks in networked control systems (Hespanhol et al., 2018). However, in this paper, the uncertainty is added to the input of the controller, while watermarking techniques usually add it to the output of the controller, i.e., the actuator signal.

### 1.3 Notation

Let  $x$  be a column vector in  $\mathbb{R}^n$  and  $A$  a matrix in  $\mathbb{R}^{n \times m}$ . The spectral radius of a square matrix  $A$  is  $\rho(A)$ . Further, we say  $A$  is (Schur) stable, if  $\rho(A) < 1$ . The trace of  $A$  is denoted as  $\text{tr}(A)$ . By  $B > 0$  ( $B \geq 0$ ), we mean a matrix is symmetric positive definite (semi-definite). The identity matrix of dimension  $n$  is denoted as  $I_n$ , while 0 denotes either a scalar, a vector, or a matrix with all elements equal to zero. The dimension of 0 is clear from the context. A Gaussian random variable  $x$  with mean  $\mu$  and covariance matrix  $\Sigma$  is denoted as  $x \sim \mathcal{N}(\mu, \Sigma)$ .

## 2 Problem formulation

In this section, we present the models of the plant and controller. Further, we describe the assumptions on and the goals of the attacker, which set the stage for the formulation of the problem.

### 2.1 Plant and controller model

The plant is modelled as a linear discrete-time system,

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) + w(k), \\ y(k) &= Cx(k) + v(k), \end{aligned} \quad (1)$$

where  $x(k)$  is the state of the plant in  $\mathbb{R}^{n_x}$ ,  $u(k)$  is the plant input in  $\mathbb{R}^{n_u}$ , and  $y(k)$  is the measured output in  $\mathbb{R}^{n_y}$ . Further,  $A \in \mathbb{R}^{n_x \times n_x}$  is the system matrix,  $B \in \mathbb{R}^{n_x \times n_u}$  is the input matrix, and  $C \in \mathbb{R}^{n_y \times n_x}$  is the output matrix. Here,  $w(k) \sim \mathcal{N}(0, \Sigma_w)$  is the process noise and  $v(k) \sim \mathcal{N}(0, \Sigma_v)$  is the measurement noise, where  $\Sigma_w \geq 0$  and  $\Sigma_v > 0$  are the covariance matrices

of the respective noise terms and have appropriate dimensions. The noise processes  $w(k)$  and  $v(k)$  are each independent and mutually uncorrelated. The operator uses an output-feedback controller of the form

$$\begin{aligned} x_c(k+1) &= A_c x_c(k) + B_c y(k), \\ u(k) &= C_c x_c(k) + D_c y(k), \end{aligned} \quad (2)$$

where  $x_c(k)$  is the controller's state in  $\mathbb{R}^{n_c}$ ,  $A_c \in \mathbb{R}^{n_c \times n_c}$  is the system matrix of the controller,  $B_c \in \mathbb{R}^{n_c \times n_y}$  is the input matrix of the controller,  $C_c \in \mathbb{R}^{n_u \times n_c}$  is the output matrix of the controller, and  $D_c \in \mathbb{R}^{n_u \times n_y}$  is the feedthrough matrix from the measurements to the actuator signal. This structure can represent many commonly used controllers. For example, with  $A_c = A - BK - LC$ ,  $B_c = L$ ,  $C_c = -K$ , and  $D_c = 0$ , we obtain an observer-based controller, where  $x_c(k)$  is an estimate of  $x(k)$ , and  $K$  and  $L$  represent the feedback and observer gain, respectively. The observer-based controller is, for example, used in Murguia and Ruths (2016).

The closed-loop system dynamics can be written as

$$\begin{aligned} \begin{bmatrix} x(k+1) \\ x_c(k+1) \end{bmatrix} &= \begin{bmatrix} A + BD_c C & BC_c \\ B_c C & A_c \end{bmatrix} \begin{bmatrix} x(k) \\ x_c(k) \end{bmatrix} \\ &\quad + \begin{bmatrix} w(k) + BD_c v(k) \\ B_c v(k) \end{bmatrix}. \end{aligned}$$

Introducing

$$z(k) = \begin{bmatrix} x(k) \\ x_c(k) \end{bmatrix} \text{ and } \eta'(k) = \begin{bmatrix} w(k) + BD_c v(k) \\ B_c v(k) \end{bmatrix}$$

we write the closed-loop system as

$$\begin{aligned} z(k+1) &= A'_z z(k) + \eta'(k) \\ y(k) &= C_z z(k) + v(k) = \begin{bmatrix} C & 0 \end{bmatrix} z(k) + v(k), \end{aligned} \quad (3)$$

where  $\eta'(k) \sim \mathcal{N}(0, Q')$  is the zero mean process noise of the closed-loop system with covariance matrix  $Q' \in \mathbb{R}^{(n_x+n_c) \times (n_x+n_c)}$  and  $v(k)$  is the measurement noise.

**Assumption 1** *The system is such that*

- (1)  $(A, B)$  is stabilizable,
- (2)  $(C, A)$  is detectable,
- (3)  $(A, \Sigma_w^{\frac{1}{2}})$  has no uncontrollable modes on the unit circle, and
- (4) the controller  $(A_c, B_c, C_c, D_c)$  is minimal.

The stability of  $A'_z$  depends on the controller matrices  $A_c, B_c, C_c$ , and  $D_c$ . Therefore, we need the first two

points of Assumption 1 such that the operator is able to observe and control all unstable modes in the system. The third point is needed later for the existence of the solution of a Riccati equation. To avoid unnecessary dynamics, the implementation of the controller should be its minimal realization.

**Assumption 2** *The operator has designed  $A_c, B_c, C_c$ , and  $D_c$ , such that the closed-loop system is stable, i.e.,  $\rho(A'_z) < 1$ .*

Assuming a stable closed-loop system is in line with normal operator requirements.

**Assumption 3** *The closed-loop system has reached steady state before  $k = 0$  and  $z(0) \sim \mathcal{N}(0, \Sigma_0)$ , where  $\Sigma_0 \geq 0$  is the solution to*

$$\Sigma_0 = A'_z \Sigma_0 (A'_z)^T + Q'.$$

This assumption is not restrictive, since industrial plants usually run for long periods of time, and we know that the covariance of  $z(k)$  will reach its unique steady state, since  $\rho(A'_z) < 1$  by Assumption 2.

Note that the closed-loop process noise variable  $\eta'(k)$  is correlated with the measurement noise  $v(k)$ ,

$$\begin{aligned} & \mathbb{E} \left\{ \begin{bmatrix} \eta'(k) \\ v(k) \end{bmatrix} \begin{bmatrix} \eta'(k)^T & v(k)^T \end{bmatrix} \right\} \\ &= \begin{bmatrix} \Sigma_w + BD_c \Sigma_v D_c^T B^T & BD_c \Sigma_v B_c^T & BD_c \Sigma_v \\ B_c \Sigma_v D_c^T B^T & B_c \Sigma_v B_c^T & B_c \Sigma_v \\ \Sigma_v B^T D_c^T & \Sigma_v^T B_c^T & \Sigma_v \end{bmatrix} \\ &= \begin{bmatrix} Q' & S \\ S^T & R \end{bmatrix}, \end{aligned}$$

where  $S \in \mathbb{R}^{(n_x+n_c) \times n_y}$ , and  $R \in \mathbb{R}^{n_y \times n_y}$ .

Since the  $\eta'(k)$  and  $v(k)$  are correlated, we will apply a transformation proposed in Chan et al. (1984) to obtain a system representation with uncorrelated noises.

$$\begin{aligned} z(k+1) &= A'_z z(k) + \eta'(k) - SR^{-1}(y(k) - y(k)) \\ &= A_z z(k) + \eta(k) + SR^{-1}y(k), \end{aligned}$$

where  $A_z = A'_z - SR^{-1}C_z$ ,

$$\eta(k) = \eta'(k) - SR^{-1}v(k) = \begin{bmatrix} w(k) \\ 0 \end{bmatrix},$$

$$\mathbb{E} \left\{ \begin{bmatrix} \eta(k) \\ v(k) \end{bmatrix} \begin{bmatrix} \eta(k)^T & v(k)^T \end{bmatrix} \right\} = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix},$$

and

$$Q = Q' - SR^{-1}S^T = \begin{bmatrix} \Sigma_w & 0 \\ 0 & 0 \end{bmatrix}.$$

The zero elements in  $Q$  show us that there is no process noise acting on the controller in the transformed system.

Therefore, the closed-loop dynamics we consider from now on are

$$\begin{aligned} z(k+1) &= A_z z(k) + \eta(k) + SR^{-1}y(k), \\ y(k) &= C_z z(k) + v(k). \end{aligned} \quad (4)$$

Note that even though  $\rho(A'_z) < 1$ , it is not always the case that  $\rho(A_z) < 1$ .

## 2.2 Attack model and goals

Now that we introduced the plant and controller model, we look into the attack model and the attacker's goal. We begin by introducing the assumptions made about the attacker.

**Assumption 4** *The attacker has gained access to the model  $(A, B, C, A_c, B_c, C_c, D_c)$ , the noise statistics  $(\Sigma_w, \Sigma_v)$ , the measurements  $y(k)$  for  $k \geq 0$  but not the control signals  $u(k)$  and the initial state of the system  $z(0)$ .*

Since the manipulation of control signals can lead to an immediate physical impact, we assume  $u(k)$  is better protected and therefore the attacker does not have access to it. Moreover, we set the start of the attack arbitrarily to  $k = 0$ . This can be interpreted as the point in time, from which the attacker has access to the measurements. From Assumption 3 we know that the plant and controller have been running for a long time. Therefore, the attacker does not know the state  $z(0)$  when it gains access to the sensor measurements.

**Assumption 5** *The attacker uses measurements up to time step  $k$  to estimate the controller's internal state at time step  $k+1$ .*

It is possible to use measurements up to time step  $k^* \geq k+1$  to estimate the controller's state at time step  $k+1$ . However, if the attacker wants to launch a false-data injection attack at time step  $k+1$ , this estimate needs to be available already.

The *goal* of the attacker is to obtain an estimate  $\hat{x}_c(k)$ , such that this estimate perfectly tracks the controller

state  $x_c(k)$  as  $k$  grows large. This can be either a first step in a larger attack scheme or a way to gain some insight in the controller's internal state. The goal can be formulated as the following problem.

**Problem 1** *Estimate  $x_c(k)$  such that the estimation error is unbiased, i.e.,  $\mathbb{E}\{x_c(k) - \hat{x}_c(k)\} = 0$ , and its covariance matrix  $\Sigma_c(k)$  approaches zero, i.e.,*

$$\lim_{k \rightarrow \infty} \Sigma_c(k) = 0$$

for a given  $\Sigma_c(0) \geq 0$ .

An estimation error covariance matrix  $\Sigma_c(k)$  that approaches zero as  $k$  grows large means the estimate converges to the true value in mean square (and thus also in probability).

Note that the controller has a minimal realization (see Assumption 1), and having access to both actuator and measurement signals would mean we can always perfectly estimate its state using a standard observer involving only the controller model.

In Section 3 we characterize for which systems the controller's confidentiality can be broken (Problem 1), and in Section 4 we discuss possible defence mechanisms.

### 3 Estimating the controller's state $x_c(k)$

In this section, we investigate when a solution to Problem 1 exists. It may seem obvious that an attacker according to Assumption 4 is without any doubt able to estimate the controller's state  $x_c(k)$  perfectly. However, we show in the following that this is not always the case. First, we present the optimal attack strategy to estimate  $x_c(k)$  and then state conditions for the convergence of  $\Sigma_c(k)$  to zero. Following this, we look into non-optimal strategies to solve Problem 1.

#### 3.1 Optimal attack strategy

To obtain the optimal attack strategy, we start by investigating the conditional probability of the closed-loop system state  $z(k+1)$  given all measurements up to time step  $k$ . Due to the presence of the process noise,  $\eta(k)$ , and measurement noise,  $v(k)$ , we know that  $z(k+1)$  is a random variable. Since (4) is a linear system with Gaussian noise, we know that  $z(k+1)$  given the measurements up to time step  $k$  is also a Gaussian random variable (Anderson and Moore, 1979). Let  $\{y(i)\}_{i=0}^k$  be the sequence  $\{y(0), \dots, y(k)\}$ , then the conditional probability distribution of  $z(k+1)$  given  $\{y(i)\}_{i=0}^k$  is

$$z(k+1|\{y(i)\}_{i=0}^k) \sim \mathcal{N}(\hat{z}(k+1), \Sigma_z(k+1)),$$

where

$$\hat{z}(k+1) = A_z \hat{z}(k) + SR^{-1}y(k) + L_z(k)(y(k) - C_z \hat{z}(k)) \quad (5)$$

is the conditional mean of  $z(k+1)$  with  $L_z(k) = (A_z \Sigma_z(k) C_z^T)(C_z \Sigma_z(k) C_z^T + R)^{-1}$ ,  $\hat{z}(0) = \mathbb{E}\{z(0)\} = 0$ , and

$$\begin{aligned} \Sigma_z(k+1) &= A_z \Sigma_z(k) A_z^T + Q \\ &\quad - (A_z \Sigma_z(k) C_z^T)(C_z \Sigma_z(k) C_z^T + R)^{-1} (A_z \Sigma_z(k) C_z^T)^T \end{aligned} \quad (6)$$

is the conditional covariance matrix. Its initial condition is  $\Sigma_z(0) = \Sigma_0$ , which is given in Assumption 3.

The optimal estimator for  $z(k)$  given  $\{y(i)\}_{i=0}^k$  is the Kalman filter (Anderson and Moore, 1979). It is optimal in the sense that it minimizes the mean square error. Therefore, the *optimal* attack strategy to estimate  $x_c(k)$  is a time-varying Kalman filter, which uses  $\hat{z}(k)$  in (5) as the estimate of  $z(k)$ . The goal of the attacker is to have an estimate  $\hat{z}(k)$  of the closed-loop system's state such that  $\begin{bmatrix} 0 & I_{n_c} \end{bmatrix} \hat{z}(k) \rightarrow x_c(k)$  as  $k \rightarrow \infty$ .

Instead of directly analysing  $\hat{z}(k)$ , we introduce the estimation error  $e_z(k) = z(k) - \hat{z}(k)$  that has the dynamics

$$e_z(k+1) = (A_z - L_z(k)C_z)e_z(k) + \eta(k) + L_z(k)v(k).$$

and covariance matrix

$$\mathbb{E}\{e_z(k+1)e_z(k+1)^T\} = \Sigma_z(k+1).$$

A Kalman filter is an unbiased estimator, which means that  $\mathbb{E}\{z(k)\} = \hat{z}(k)$ , or, differently formulated,  $\mathbb{E}\{e_z(k)\} = 0$ . Hence, Problem 1 is solved if, for  $\Sigma_z(0) = \Sigma_0$ , the attacker's Kalman filter fulfils

$$\lim_{k \rightarrow \infty} \Sigma_z(k) = \begin{bmatrix} P & 0 \\ 0 & 0 \end{bmatrix}, \quad (7)$$

where  $P \geq 0$ . Note that  $\Sigma_0$  can be calculated by the attacker because of its model knowledge by Assumption 4.

#### 3.2 Asymptotic convergence to $\Sigma_c(k) = 0$

Let us now investigate when the optimal attack strategy solves Problem 1. Here, we present necessary and sufficient conditions for the covariance matrix  $\Sigma_c(k)$  to converge to zero. Recall this is equivalent to saying that (7) is fulfilled.

Before we present our convergence results, note that a steady state solution to (6) satisfies the algebraic Riccati

equation (ARE)

$$\Sigma_\infty = A_z \Sigma_\infty A_z^T + Q - (A_z \Sigma_\infty C_z^T)(C_z \Sigma_\infty C_z^T + R)^{-1}(A_z \Sigma_\infty C_z^T)^T, \quad (8)$$

where  $L_\infty = (A_z \Sigma_\infty C_z^T)(C_z \Sigma_\infty C_z^T + R)^{-1}$  is the steady state Kalman gain.

**Definition 1 (Definition 3.1 (Chan et al., 1984))**

A real symmetric nonnegative definite solution  $\Sigma_\infty$  to (8) is called a strong solution if  $\rho(A_z - L_\infty C_z) \leq 1$ . The strong solution is called a stabilizing solution if  $\rho(A_z - L_\infty C_z) < 1$ .

The following lemma from de Souza et al. (1986) will be useful in the following discussion.

**Lemma 1 (Theorem 3.2 (de Souza et al., 1986))**

Let  $G^T G = Q$ ,

- (1) the strong solution of the ARE exists and is unique if and only if  $(C_z, A_z)$  is detectable;
- (2) the strong solution is the only nonnegative definite solution of the ARE if and only if  $(C_z, A_z)$  is detectable and  $(A_z, G)$  has no uncontrollable modes outside the unit circle;
- (3) the strong solution coincides with the stabilizing solution if and only if  $(C_z, A_z)$  is detectable and  $(A_z, G)$  has no uncontrollable modes on the unit circle;
- (4) the stabilizing solution is positive definite if and only if  $(C_z, A_z)$  is detectable and  $(A_z, G)$  has no uncontrollable modes inside, or on the unit circle.

Let us begin by showing that a solution to (8) of the form in (7) exists.

**Proposition 1** A solution of the algebraic Riccati equation (8) is given by

$$\Sigma_\infty = \begin{bmatrix} P & 0 \\ 0 & 0 \end{bmatrix},$$

where  $P \geq 0$  is the unique solution of the ARE

$$P = APA^T + \Sigma_w - APC^T(CPC^T + \Sigma_v)^{-1}CPA^T.$$

**PROOF.** Let us first determine

$$A_z = A'_z - SR^{-1}C_z = \begin{bmatrix} A & BC_c \\ 0 & A_c \end{bmatrix}.$$

After algebraic computations we obtain

$$A_z \Sigma_\infty A_z^T + Q = \begin{bmatrix} APA^T + \Sigma_w & 0 \\ 0 & 0 \end{bmatrix}, \quad A_z \Sigma_\infty C_z^T = \begin{bmatrix} APC^T \\ 0 \end{bmatrix},$$

and  $C_z \Sigma_\infty C_z^T + R = CPC^T + \Sigma_v$  such that

$$(A_z \Sigma_\infty C_z^T)(C_z \Sigma_\infty C_z^T + R)^{-1}(A_z \Sigma_\infty C_z^T)^T = \begin{bmatrix} APC^T(CPC^T + \Sigma_v)^{-1}CPA^T & 0 \\ 0 & 0 \end{bmatrix}.$$

This leads to

$$\Sigma_\infty = \begin{bmatrix} APA^T + \Sigma_w - APC^T(CPC^T + \Sigma_v)^{-1}CPA^T & 0 \\ 0 & 0 \end{bmatrix}.$$

For  $\Sigma_\infty$  to be a solution of (8) we require

$$P = APA^T + \Sigma_w - APC^T(CPC^T + \Sigma_v)^{-1}CPA^T. \quad (9)$$

Note that (9) by itself is an algebraic Riccati equation. It is actually the algebraic Riccati equation an operator would obtain when it is designing a time-invariant Kalman filter. Due to detectability of  $(C, A)$  (Assumption 1), there exists a unique strong solution  $P \geq 0$  for (9) (Lemma 1). Hence,  $\Sigma_\infty$  is a solution of (8).

Now that we proved that  $\Sigma_\infty$  is indeed a solution to the algebraic Riccati equation, we need to show under which conditions  $\Sigma_z(k)$  converges to  $\Sigma_\infty$  for the initial condition  $\Sigma_0$ .

**Lemma 2** The unique strong solution of the ARE (8) is  $\Sigma_\infty$  if and only if  $\rho(A_c) \leq 1$ .

**PROOF.** Due to the first statement in Lemma 1, the strong solution is unique and exists if and only if  $(C_z, A_z)$  is detectable. From the stability of  $A'_z = A_z + SR^{-1}C_z$ , it follows that  $(C_z, A_z)$  is detectable. Hence, the strong solution will be unique. Further, if  $\rho(A_z - L_\infty C_z) \leq 1$  for

$$L_\infty = (A_z \Sigma_\infty C_z^T)(C_z \Sigma_\infty C_z^T + R)^{-1} = \begin{bmatrix} APC^T(CPC^T + \Sigma_v)^{-1} \\ 0 \end{bmatrix} = \begin{bmatrix} \bar{L} \\ 0 \end{bmatrix},$$

then  $\Sigma_\infty$  is a strong solution. Let us now look at the eigenvalues of  $A_z - L_\infty C_z$ , which are determined by the

eigenvalues of  $A - \bar{L}C$  and  $A_c$ , because

$$A_z - L_\infty C_z = \begin{bmatrix} A - \bar{L}C & BC_c \\ 0 & A_c \end{bmatrix}.$$

Due to the detectability of  $(C, A)$  (Assumption 1), the first statement of Lemma 1 shows us that  $P$  is a strong solution of (9), such that  $\rho(A - \bar{L}C) \leq 1$ . Therefore,  $\rho(A_z - L_\infty C_z) \leq 1$ , i.e.,  $\Sigma_\infty$  is the unique strong solution, if and only if  $\rho(A_c) \leq 1$ .

**Theorem 1** *The covariance matrix  $\Sigma_z(k)$  converges to the attacker's desired covariance matrix  $\Sigma_\infty$  for the initial condition  $\Sigma_0$  if and only if  $\rho(A_c) \leq 1$ .*

**PROOF.** By Lemma 2,  $\Sigma_\infty$  is the unique strong solution of (8) if and only if  $\rho(A_c) \leq 1$ . Theorem 4.2 in de Souza et al. (1986) states that subject to  $\Sigma_0 - \Sigma_\infty \geq 0$  the covariance matrix  $\Sigma_z(k)$  will converge to the strong solution  $\Sigma_\infty$  if and only if  $(C_z, A_z)$  is detectable. That  $(C_z, A_z)$  is detectable is shown in the proof of Lemma 2. Let us now show that  $\Sigma_0 - \Sigma_\infty \geq 0$ . If we use the system representation with correlated noise processes (3), the ARE for  $\Sigma_\infty$ , according to Anderson and Moore (1979), is

$$\begin{aligned} \Sigma_\infty &= A'_z \Sigma_\infty (A'_z)^T + Q' \\ &- (A'_z \Sigma_\infty C_z^T + S)(C_z \Sigma_\infty C_z^T + R)^{-1} (A'_z \Sigma_\infty C_z^T + S)^T. \end{aligned} \quad (10)$$

Subtracting (10) from the Lyapunov equation for  $\Sigma_0$  in Assumption 3 leads to

$$\begin{aligned} \Sigma_0 - \Sigma_\infty &= A'_z (\Sigma_0 - \Sigma_\infty) (A'_z)^T \\ &+ (A'_z \Sigma_\infty C_z^T + S)(C_z \Sigma_\infty C_z^T + R)^{-1} (A'_z \Sigma_\infty C_z^T + S)^T. \end{aligned}$$

This is also a Lyapunov equation with a unique solution since  $\rho(A'_z) < 1$  (Assumption 2). Further, we observe that

$$(A'_z \Sigma_\infty C_z^T + S)(C_z \Sigma_\infty C_z^T + R)^{-1} (A'_z \Sigma_\infty C_z^T + S)^T \geq 0,$$

because  $\Sigma_\infty \geq 0$ . Therefore, we know that  $\Sigma_0 - \Sigma_\infty \geq 0$ . Hence, with initial condition  $\Sigma_0$

$$\lim_{k \rightarrow \infty} \Sigma_z(k) = \Sigma_\infty$$

if and only if  $\rho(A_c) \leq 1$ .

**Corollary 1** *Problem 1 is solvable if and only if  $\rho(A_c) \leq 1$ .*

Note that since the attacker uses a Kalman filter, it does not only obtain a perfect estimate of  $x_c(k)$  but also an optimal estimate of  $x(k)$ .

Theorem 1 shows that the covariance matrix converges to the attacker's desired strong solution, but not how fast the convergence is. Therefore, we will now investigate the conditions for an exponential convergence rate.

**Proposition 2** *Subject to  $\Sigma_0 > 0$ , the covariance matrix  $\Sigma_z(k)$  converges exponentially fast to  $\Sigma_\infty$  if and only if  $\rho(A_c) < 1$ .*

**PROOF.** Theorem 4.1 in de Souza et al. (1986) shows us that subject to  $\Sigma_0 > 0$  the covariance matrix  $\Sigma_z(k)$  converges exponentially fast to the stabilizing solution if and only if  $(C_z, A_z)$  is detectable and  $(A_z, G)$  has no uncontrollable modes on the unit circle. We already showed that  $(C_z, A_z)$  is detectable, therefore we look at the controllable modes of  $(A_z, G)$  now. Recall that  $GG^T = Q$  such that

$$G = \begin{bmatrix} \Sigma_w^{\frac{1}{2}} & 0 \\ 0 & 0 \end{bmatrix}.$$

For  $(A_z, G)$  to have no uncontrollable modes on the unit circle we need  $A_c$  to have no eigenvalues on the unit circle, because we cannot control the eigenvalues of  $A_c$  with  $G$ , and due to Assumption 1  $(A, \Sigma_w^{\frac{1}{2}})$  has no uncontrollable modes on the unit circle. We showed in Lemma 2 that  $\Sigma_\infty$  is a strong solution to the ARE if and only if  $\rho(A_c) \leq 1$ . Hence, subject to  $\Sigma_0 > 0$  the covariance matrix  $\Sigma_z(k)$  converges exponentially fast to  $\Sigma_\infty$  if and only if  $\rho(A_c) < 1$ .

This shows us that if  $\Sigma_0 > 0$  and the operator uses a stable controller, i.e.,  $\rho(A_c) < 1$ , the covariance matrix of the attacker's time-varying Kalman filter will converge exponentially fast to  $\Sigma_\infty$ . Hence, the attacker is able to obtain a perfect estimate of  $x_c(k)$  exponentially fast.

### 3.3 Breaking confidentiality of $x_c(k)$ using non-optimal observers

Previously, we have shown under which conditions the attacker is able to get a perfect estimate of the controller state  $x_c(k)$  when a time-varying Kalman filter is used. The time-varying Kalman filter is the optimal filter for linear systems with Gaussian noise. One may wonder whether or not the attacker is able to perfectly estimate  $x_c(k)$ , when the attacker uses a non-optimal observer. Here, we investigate a time-invariant observer of the form

$$\hat{z}(k+1) = A_z \hat{z}(k) + SR^{-1}y(k) + L_z(y(k) - C_z \hat{z}(k)), \quad (11)$$

with  $\hat{z}(0) = 0$ , where  $L_z$  is the attacker's constant observer gain. As before, instead of looking at  $\hat{z}(k)$ , we

analyse the error dynamics given by

$$e_z(k+1) = (A_z - L_z C_z) e_z(k) + \eta(k) + L_z v(k).$$

with  $\mathbb{E}\{e_z(k)\} = 0$  for all  $k \geq 0$ , covariance matrix  $\mathbb{E}\{e_z(k)e_z(k)^T\} = \Sigma_z(k)$  and  $\Sigma_z(0) \geq 0$ .

The following theorem classifies all gains  $L_z$  of a non-optimal observer such that Problem 1 is solved.

**Theorem 2** *For any  $\Sigma_z(0) \geq 0$ ,*

$$\lim_{k \rightarrow \infty} \Sigma_z(k) = \tilde{\Sigma}_\infty = \begin{bmatrix} \tilde{P} & 0 \\ 0 & 0 \end{bmatrix},$$

*if and only if  $\rho(A_c) < 1$ ,  $L_z = [L_1^T \ 0^T]^T$  and  $L_1 \in \mathbb{R}^{n_x \times n_y}$  is chosen such that  $\rho(A - L_1 C) < 1$ . Here,  $\tilde{P}$  is the unique solution to*

$$\tilde{P} = (A - L_1 C) \tilde{P} (A - L_1 C)^T + \Sigma_w + L_1 \Sigma_v L_1^T,$$

*and  $\tilde{P} - P \geq 0$ , where  $P$  is the unique solution to (9).*

**PROOF.** With  $L_z = [L_1^T \ L_2^T]^T$  the error dynamics are

$$e_z(k+1) = \begin{bmatrix} A - L_1 C & B C_c \\ -L_2 C & A_c \end{bmatrix} e_z(k) + \begin{bmatrix} w(k) - L_1 v(k) \\ L_2 v(k) \end{bmatrix}.$$

The error covariance matrix evolves as

$$\Sigma_z(k+1) = (A_z - L_z C_z) \Sigma_z(k) (A_z - L_z C_z)^T + \begin{bmatrix} \Sigma_w + L_1 \Sigma_v L_1^T & L_1 \Sigma_v L_2^T \\ L_2 \Sigma_v L_1^T & L_2 \Sigma_v L_2^T \end{bmatrix}. \quad (12)$$

Now we show that  $\tilde{\Sigma}_\infty$  is the steady state solution of (12) if and only if  $L_2 = 0$ . First, we observe that if  $L_2 = 0$  then  $\tilde{\Sigma}_\infty$  is a steady state solution of (12), where  $\tilde{P}$  is the solution to the Lyapunov equation

$$\tilde{P} = (A - L_1 C) \tilde{P} (A - L_1 C)^T + \Sigma_w + L_1 \Sigma_v L_1^T.$$

Note that  $\tilde{P} \geq 0$  exists and is unique if  $\rho(A - L_1 C) < 1$ . Second, if  $\tilde{\Sigma}_\infty$  is a steady state solution of (12) the equations

$$\begin{aligned} \tilde{P} &= (A - L_1 C) \tilde{P} (A - L_1 C)^T + \Sigma_w + L_1 \Sigma_v L_1^T, \\ 0 &= L_2 (\Sigma_v L_1^T - C \tilde{P} (A - L_1 C)^T), \text{ and} \\ 0 &= L_2 (C \tilde{P} C^T + \Sigma_v) L_2^T \end{aligned}$$

are fulfilled. The last equation is only fulfilled if  $L_2 = 0$ , since  $\Sigma_v$  is positive definite. This simultaneously fulfils the second equation. The first equation recovers the

Lyapunov equation for  $\tilde{P}$ . Therefore, if  $\tilde{\Sigma}_\infty$  is a steady state solution of (12) then  $L_2 = 0$ . Hence, (12) has  $\tilde{\Sigma}_\infty$  as a steady state solution if and only if  $L_2 = 0$ . Let us now look at the convergence of (12) to  $\tilde{\Sigma}_\infty$ . For any  $\Sigma_z(0) \geq 0$ , the error covariance matrix converges to  $\tilde{\Sigma}_\infty$  if and only if  $\rho(A_z - L_z C_z) < 1$ . With  $L_2 = 0$ , the stability of  $A_z - L_z C_z$  is guaranteed when both  $\rho(A_c) < 1$  and  $\rho(A - L_1 C) < 1$ . Due to detectability of  $(C, A)$  in Assumption 1 such a stabilizing  $L_1$  exists. Therefore, (12) converges to  $\tilde{\Sigma}_\infty$  for any  $\Sigma_z(0) \geq 0$ , if and only if  $L_2 = 0$ ,  $\rho(A - L_1 C) < 1$ , and  $\rho(A_c) < 1$ . Further,  $\rho(A_z - L_z C_z) < 1$  also makes  $\tilde{\Sigma}_\infty$  the unique steady state solution of (12). Since the Kalman filter is the best linear estimator, we know that  $\tilde{P} - P \geq 0$  and  $\tilde{P} = P$  if  $L_1 = APC^T(CPC^T + \Sigma_v)^{-1}$  (Anderson and Moore, 1979). This choice of  $L_1$  turns the Lyapunov equation of  $\tilde{P}$  into (9).

Theorem 2 shows us that the attacker is able to use the non-optimal observer (11) to solve Problem 1, if and only if the controller is stable.

**Corollary 2** *Problem 1 is solvable with a non-optimal observer of the form (11) if and only if  $\rho(A_c) < 1$ .*

According to Theorem 2, the attacker does not need to know the noise statistics  $\Sigma_w$  and  $\Sigma_v$  for the design of  $L_1$  to estimate  $x_c(k)$  perfectly, as long as  $L_1$  is stabilizing. Hence, the attacker's required knowledge to solve Problem 1 is reduced when the operator uses a stable controller. Further, the attacker has a smaller computational burden when a time-invariant observer is used.

## 4 Defence mechanisms

We presented under which conditions Problem 1 is solvable both with optimal and non-optimal strategies. Therefore, we investigate now how to prevent the attacker from estimating  $x_c(k)$  perfectly, i.e., make Problem 1 unsolvable. We present a defence mechanism and discuss why an unstable controller is only in certain cases a good defence mechanism.

### 4.1 Injecting noise on the controller side

As previously shown, an attacker under Assumption 4 will be able to predict the controller state perfectly for  $\rho(A_c) \leq 1$ . We observe that the controller dynamics in (2) contain no uncertainty for the attacker when  $y(k)$  is known. Therefore, an approach for defence is to introduce uncertainty in the form of an additional noise term on the controller side.

The additional noise term  $\nu(k)$  has a zero mean Gaussian distribution with a positive semi-definite covariance



matrix  $\Sigma_\nu \in \mathbb{R}^{n_c \times n_c}$ . Further,  $\nu(k)$  is independent and identically distributed over time and also independent of  $w(k)$ ,  $v(k)$ , and  $z(0)$ . The controller state with the additional noise term follows the dynamics

$$x_c(k+1) = A_c x_c(k) + B_c y(k) + \nu(k).$$

Here,  $\nu(k)$  can be interpreted as process noise of the controller.

**Assumption 6** *The attacker knows the covariance matrix  $\Sigma_\nu$  of the additional noise in the controller.*

This assumption is in the spirit of Assumption 4, since the attacker has full model knowledge and knows the noise statistics of both  $w(k)$  and  $v(k)$ .

This changes the process noise of the closed-loop system (4) from  $\eta(k)$  to  $\tilde{\eta}(k) = [w(k)^T \ \nu(k)^T]^T$  such that

$$\mathbb{E} \left\{ \begin{bmatrix} \tilde{\eta}(k) \\ v(k) \end{bmatrix} \begin{bmatrix} \tilde{\eta}(k)^T & v(k)^T \end{bmatrix} \right\} = \left[ \begin{array}{cc|c} \Sigma_w & 0 & 0 \\ 0 & \Sigma_\nu & 0 \\ \hline 0 & 0 & \Sigma_v \end{array} \right] = \left[ \begin{array}{c|c} \tilde{Q} & 0 \\ \hline 0 & R \end{array} \right].$$

The following proposition shows that with  $\nu(k)$ , the attacker's desired covariance matrix  $\Sigma_\infty$  is not a steady state solution of (8) any more.

**Proposition 3** *The algebraic Riccati equation (8) with  $Q = \tilde{Q}$  does not have  $\Sigma_\infty$  as a steady state solution.*

**PROOF.** With  $\Sigma_z(k) = \Sigma_\infty$  and  $Q = \tilde{Q}$  we obtain

$$A_z \Sigma_\infty A_z^T + \tilde{Q} = \begin{bmatrix} APA^T + \Sigma_w & 0 \\ 0 & \Sigma_\nu \end{bmatrix},$$

and using this in the Riccati equation (8) leads to

$$\Sigma_\infty = \begin{bmatrix} APA^T + \Sigma_w - APC^T(CPC^T + \Sigma_v)^{-1}CPA^T & 0 \\ 0 & \Sigma_\nu \end{bmatrix}.$$

For  $\Sigma_\infty$  to be a solution of (8) we need both

$$P = APA^T + \Sigma_w - APC^T(CPC^T + \Sigma_v)^{-1}CPA^T,$$

which, as shown previously, exists, and  $\Sigma_\nu = 0$ .

Since we assume  $\Sigma_\nu \neq 0$ ,  $\Sigma_\infty$  is not a solution of (8) any more.

Here, we see that the attacker will not be able to perfectly estimate the controller's state if we use this additional

noise on the controller side even if the attacker knows the noise properties.

Injecting  $\nu(k)$  does not only lead to  $\lim_{k \rightarrow \infty} \Sigma_z(k) = \tilde{\Sigma}_\infty \neq \Sigma_\infty$  as shown in Proposition 3 but also changes the steady state covariance matrix of the closed-loop system from  $\Sigma_0$  (see Assumption 3) to  $\tilde{\Sigma}_0$ . The change in the covariance matrix,  $\Delta\Sigma_0 = \tilde{\Sigma}_0 - \Sigma_0$ , is given by

$$\Delta\Sigma_0 = A'_z \Delta\Sigma_0 (A'_z)^T + \underbrace{\begin{bmatrix} 0 & 0 \\ 0 & \Sigma_\nu \end{bmatrix}}_{=\Delta Q}. \quad (13)$$

Furthermore, we will quantify the performance degradation in the closed-loop system (3) as  $\text{tr}(\Delta\Sigma_0)$ , which represents the increase in total variation for the closed-loop system state.

Therefore, we formulate the following convex optimization problem to determine  $\Sigma_\nu$  subject to an upper bound  $\gamma_p > 0$  on the performance degradation.

**Proposition 4** *The noise injection covariance  $\Sigma_\nu$  that maximizes the controller confidentiality while keeping the performance degradation below a threshold,  $\gamma_p > 0$ , is an optimal solution to the convex program,*

$$\begin{aligned} & \max_{\Sigma_\nu, \tilde{\Sigma}_\infty} \text{tr}(\tilde{\Sigma}_\infty) \\ & \text{s.t.} \quad \begin{cases} \begin{bmatrix} A_z \tilde{\Sigma}_\infty A_z^T + Q + \Delta Q - \tilde{\Sigma}_\infty & A_z \tilde{\Sigma}_\infty C_z^T \\ (A_z \tilde{\Sigma}_\infty C_z^T)^T & C_z \tilde{\Sigma}_\infty C_z^T + R \end{bmatrix} \geq 0, \\ \Delta\Sigma_0 = A'_z \Delta\Sigma_0 (A'_z)^T + \Delta Q \\ \text{tr}(\Delta\Sigma_0) \leq \gamma_p, \\ \Sigma_\nu \geq 0, \quad \tilde{\Sigma}_\infty \geq 0, \end{cases} \end{aligned} \quad (14)$$

where  $\Delta Q$  is given in (13).

**PROOF.** First note that both the objective and the constraints are convex in  $\Sigma_\nu$  and  $\tilde{\Sigma}_\infty$ , which makes the optimization problem a convex semi-definite program and that problem (14) is feasible, since  $\Sigma_\nu = 0$  and  $\tilde{\Sigma}_\infty = \Sigma_\infty$  fulfil the constraints.

Next, the objective together with the first constraint guarantee that  $\tilde{\Sigma}_\infty$  is the solution to the algebraic Riccati equation (8), since the solution to (8) is the maximal solution to the algebraic Riccati inequality (Ran and Vreugdenhil, 1988).

Last, the second and third constraint impose the limitation on the allowed performance degradation, while the last two constraints enforce that the covariance matrices are positive semi-definite.

**Remark 1** If a certain noise level in the controller is desired, the constraint  $\Sigma_\nu \geq 0$  can be replaced by  $\Sigma_\nu \geq \gamma_c I_{n_c}$ , where  $\gamma_c > 0$ . However, this tighter constraint can return an optimal  $\tilde{\Sigma}_\infty$  with a smaller trace than in the case with  $\gamma_c = 0$ . Furthermore, this constraint can make the problem infeasible, since a large  $\gamma_c$  can interfere with the constraint on the performance bound.

**Remark 2** The approach of adding some additional noise to the system is quite similar to the watermarking approach used, for example, in Mo et al. (2015). The difference is that here the noise is added to the controller input, while in watermarking the noise is typically added to the output of the controller. Therefore, these results show that if we position the watermarking noise at a different position we get the additional benefit of the attacker not being able to estimate the state of the controller perfectly.

#### 4.2 An unstable controller as defence

As shown before, Problem 1 is not solvable if and only if  $\rho(A_c) > 1$ . Hence, designing the controller  $(A_c, B_c, C_c, D_c)$  such that  $\rho(A'_z) < 1$  and  $\rho(A_c) > 1$  leads to a successful defence against the discussed disclosure attack.

This implies that there are plants which have an inherent protection against the sensor attack. For example, all plants that are *not* strongly stabilizable, i.e., plants that cannot be stabilized with a stable controller (Doyle et al., 2013), have an inherent protection against the estimation of the controller's state by the attacker. Further, there are also control strategies that give an inherent protection to the closed-loop system. Disturbance accommodation control (Johnson, 1971), where the controller tries to estimate a persistent disturbance, is one example of these control strategies.

If a plant can be stabilized by using a stable controller, i.e., a strongly stabilizing plant, using an unstable controller instead comes with several issues. A fundamental limitation is that the integral of the log sensitivity function is zero for a stable open-loop system. If the open-loop system has unstable poles the integral is equal to a constant positive value that depends on the unstable poles of the open-loop system and their directions for a multivariable discrete-time system (Chen and Nett, 1995). As Stein (2003) shows with real world examples, it can have dire consequences if this fundamental limitation is not taken into account properly. Hence, due to these fundamental limitation the introduction of unstable poles in the controller is not desirable. Another issue of unstable controllers is that an unstable controller leads to an unstable open-loop system, if the feedback loop is interrupted.

Therefore, using an unstable controller for a strongly stabilizing plant is not recommended, but is an appropriate

defence mechanism if an unstable controller is needed to stabilize the plant.

## 5 Simulations

In this section, we verify our results with simulations for a three-tank system. After stating the model of the three-tank system, we first show the effect of stable and unstable controllers on the attacker's estimate of the controller's state. Later, we verify that the additional noise prevents the attacker from estimating the controller's state perfectly.

### 5.1 The three-tank system

For the simulation of the closed-loop system estimation by the attacker we look at the following continuous-time three-tank system

$$\begin{aligned} \dot{x}(t) &= \begin{bmatrix} -2 & 2 & 0 \\ 2 & -4 & 2 \\ 0 & 2 & -3 \end{bmatrix} x(t) + \begin{bmatrix} 0.5 & 0 \\ 0 & 0 \\ 0 & 0.5 \end{bmatrix} u(t) + w(t), \\ y(t) &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x(t) + v(t). \end{aligned}$$

By discretizing the continuous-time system with a sampling period of  $T_s = 0.5$  s we obtain  $A$ ,  $B$ , and  $C$ . We assume that  $w(k) \sim \mathcal{N}(0, I_3)$  and  $v(k) \sim \mathcal{N}(0, 0.1I_2)$ .

### 5.2 Stable and unstable controllers

Now that the system matrices are defined we are going to verify that the controller's stability influences the estimates of the controller's state by the attacker. We consider an observer-based feedback controller

$$\begin{aligned} x_c(k+1) &= (A - BK_i - LC)x_c(k) + Ly(k) \\ u(k) &= -K_i x_c(k) \end{aligned}$$

where  $L$  is the observer gain and  $K_i$  is the controller gain. The closed-loop system matrix is then

$$A'_{z,i} = \begin{bmatrix} A & -BK_i \\ LC & A - BK_i - LC \end{bmatrix}.$$

According to Assumption 2,  $\rho(A'_{z,i}) < 1$ , which means that  $K_i$  and  $L$  are designed such that  $\rho(A - BK_i) < 1$  and  $\rho(A - LC) < 1$ . The matrix  $L$  is designed via pole placement to place the eigenvalues of  $A - LC$  at 0.1, 0.2, and 0.3. Therefore, the error dynamics of the observer used in the controller are stable. In the following, we design three different  $K_i$  such that  $\rho(A - BK_i) < 1$ .

The first controller  $K_S$  places the poles of  $A - BK_S$  at 0.4, 0.5, and 0.6. This first controller results in stable controller dynamics  $A - BK_S - LC$  with  $\rho(A - BK_S - LC) = 0.4167$ .

The second controller,  $K_U$ , is unstable, i.e.,  $\rho(A - BK_U - LC) > 1$ , but has no modes on the unit circle. We determine  $K_U$ , such that  $\rho(A - BK_U) < 1$  and  $A - BK_U - LC$  has an eigenvalue at 1.5. The controller we obtain is

$$K_U = \begin{bmatrix} 0.5530 & 1.9589 & 1.2225 \\ 1.8414 & 27.0785 & -12.9349 \end{bmatrix}$$

and it places the eigenvalues of  $A - BK_U - LC$  at 1.5,  $-0.5175$ , and  $-0.1066$  and the eigenvalues of  $A - BK_U$  at  $0.6275$ ,  $0.4272 + j0.6456$ , and  $0.4272 - j0.6456$ .

For the design of the third controller,  $K_I$ , we place two eigenvalues inside the unit circle and one at 1, such that  $\rho(A - BK_I - LC) = 1$ , while guaranteeing that  $\rho(A - BK_I) < 1$ . We obtain

$$K_I = \begin{bmatrix} 3.0988 & -6.0472 & 2.3966 \\ 4.0471 & 10.8175 & -4.4516 \end{bmatrix},$$

which places the eigenvalues of  $A - BK_I - LC$  at 1,  $-0.2227$ , and  $-0.3693$  and the eigenvalues of  $A - BK_I$  at  $-0.2669$ ,  $0.6405 + j0.5942$ , and  $0.6405 - j0.5942$ .

For the first two controllers, the attacker designs a time-invariant Kalman filter with gain  $L_z^i$  and steady state error covariance matrix  $\Sigma_\infty^i = \lim_{k \rightarrow \infty} \Sigma^i(k)$ , where  $i \in \{S, U\}$ . The attacker's time-invariant Kalman filter design leads to an observer gain  $L_z^S$  for the closed-loop system, which matches our results in Theorem 2. Since  $K_U$  leads to an unstable controller, we know according to Corollary 2 that no time-invariant observer exists that solves Problem 1. Further, Corollary 1 shows that even if the attacker would use a time-varying Kalman filter, Problem 1 is not solvable.

For the closed-loop system with  $K_I$ , the attacker needs to use a time-varying Kalman filter to obtain a perfect estimate of  $x_c(k)$ . The error covariance matrix in this case will converge to the same as in the case with  $K_S$ .

Now that we designed the Kalman filters for each of the three closed-loop systems, let us look at the estimation error  $e_z(k) = z(k) - \hat{z}(k) \in \mathbb{R}^6$ . Here, we are only interested in the last three elements of  $e_z(k)$ , because they represent the estimation error of the controller's state. The  $j$ th element of  $e_z(k)$  is denoted by  $e_{z,j}(k)$ , where  $j \in \{1, \dots, 6\}$ . Figure 2 shows that in case of a stable controller  $K_S$  the estimation error converges quickly to zero and the attacker obtains a perfect estimate of the controller's state. However, if we use an unstable controller  $K_U$  the estimation error remains noisy and the

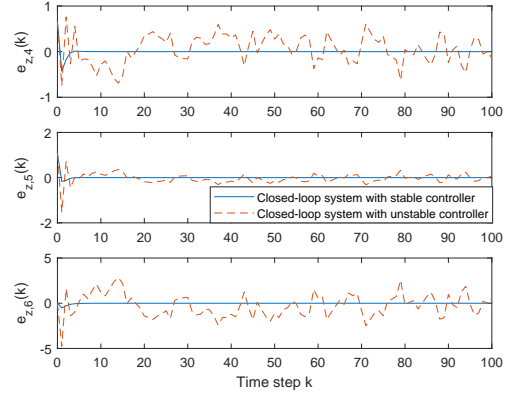


Fig. 2. Comparison of the estimation error trajectories for the stable and unstable controller,  $K_S$  and  $K_U$  respectively

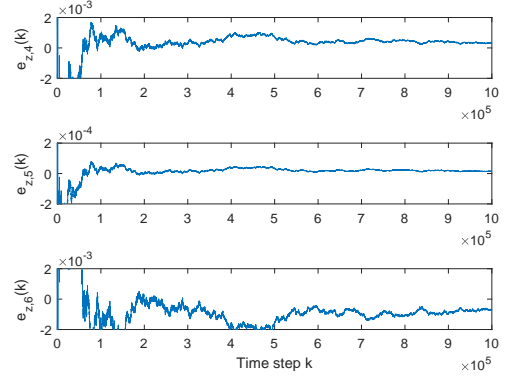


Fig. 3. Estimation error of the controller's state when the controller has a pole on the unit circle and the attacker uses a time-varying Kalman filter

attacker is not able to obtain a perfect estimate of the controller's state. Furthermore, when  $K_I$  is used, we observe that the estimation error converges to zero, but is still not zero after a million time steps (see Figure 3). Theorem 1 only tells us that the error will converge, but we know it does not converge exponentially by Proposition 2. Although the attacker can obtain an almost perfect estimate with the time-varying Kalman filter after a million time steps, it is still not a perfect estimate. This shows us that a controller with modes on the unit circle can prevent the attacker from quickly obtaining a perfect estimate.

### 5.3 Injecting process noise for the controller

Now that we showed how the controller design affects the attacker's estimate of the controller's state, we verify that injecting noise to the input of the controller prevents the attacker from estimating  $x_c(k)$  perfectly. Further, we demonstrate how the choice of  $\gamma_p$  affects both the attacker's estimate  $\hat{x}_c(k)$ , the plant's state  $x(k)$ , and the controller state  $x_c(k)$ . We investigate two cases for

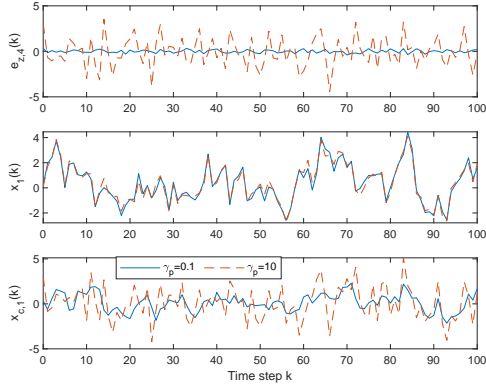


Fig. 4. The effect of the additional noise on the first elements of the estimation error of the controller's state (upper plot), the plant's state (centre plot), and the controller's state (lower plot) when a stable controller is used

the performance degradation, one with a small allowed performance degradation, i.e.,  $\gamma_p = 0.1$ , and one with a large allowed performance degradation, i.e.,  $\gamma_p = 10$ . The operator uses the stable controller  $K_S$  and the attacker uses again a time-invariant Kalman filter.

The upper plot of Figure 4 shows the trajectory of the attacker's estimation error of the first controller state. Compared to Figure 2, the estimation error exhibits noisy behaviour and the attacker is not able to obtain a perfect estimate even though the operator uses the stable controller  $K_S$ . Further, the noise around the estimation error increases the larger the allowed performance degradation is.

To see the effect of the additional noise on the closed-loop system state, we show the trajectory of the first element of the plant's state,  $x_1(k)$ , in the centre plot and the trajectory of the first element of the controller's state,  $x_{c,1}(k)$ , in the lower plot of Figure 4. We see that the state trajectory of the plant is not much more affected by the additional noise when we allow a hundred-fold larger performance degradation, while the controller state becomes noisier.

The trajectories for the other elements of  $x(k)$ ,  $x_c(k)$ , and the attacker's estimation error of the controller state behave similarly. Since the operator's objective is to control the plant optimally, this defence mechanism has the additional benefit of mostly increasing the noise in  $x_c(k)$  but not considerably in the plant's state.

Hence, the additional noise prevents the attacker from estimating the controller's state perfectly and additionally does not considerably affect the trajectory of the plant's state.

## 6 Conclusion and future work

We have shown exactly when an attacker with full model knowledge is able to perfectly estimate the internal state of an output-feedback controller by observing all measurements.

Although it seems obvious that an attacker according to our attack model can always estimate the controller's state, we gave necessary and sufficient conditions when an attacker is not able to obtain a perfect estimate. These conditions state that unstable controller dynamics prevent the attacker from obtaining a perfect estimate. Further, the attacker can use a non-optimal time-invariant observer to perfectly estimate the controller state if and only if the controller has stable dynamics. A defence mechanism has been proposed to make the controller states confidential. This mechanism prevents the attacker from obtaining a perfect estimate by adding uncertainty to the controller dynamics. This is similar to watermarking approaches proposed by other authors with the twist that the noise signal is applied to the controller input and not to its output. An unstable controller gives an inherent protection to plants that are not strongly stabilizable. However, designing such a controller introduces fundamental limitations on the sensitivity function of the closed-loop system and should only be used when an unstable controller is needed to stabilize the plant.

There are several directions of future work. It seems obvious that if the attacker has only access to a few sensors measurements, it will not be able to estimate the controller's state. However, it is interesting to investigate what happens if the attacker has access to some sensor and some actuator signals, and how many of each the attacker needs to get a perfect estimate of the controller's state. Another research direction is to investigate the robustness of the controller state estimation for cases when the attacker has less model knowledge.

## References

- B. D. O. Anderson and J. B. Moore. *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, NJ, 1979.
- A. A. Cárdenas, S. Amin, Z. Lin, Y. Huang, C. Huang, and S. Sastry. Attacks against process control systems: Risk assessment, detection, and response. In *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, ASI-ACCS '11, pages 355–366, New York, NY, USA, 2011. ACM.
- S. W. Chan, G. Goodwin, and K. S. Sin. Convergence properties of the Riccati difference equation in optimal filtering of nonstabilizable systems. *IEEE Transactions on Automatic Control*, 29(2):110–118, February 1984.

- J. Chen and C. N. Nett. Sensitivity integrals for multivariable discrete-time systems. *Automatica*, 31(8): 1113 – 1124, 1995.
- C. de Souza, M. Gevers, and G. Goodwin. Riccati equations in optimal filtering of nonstabilizable systems having singular state transition matrices. *IEEE Transactions on Automatic Control*, 31(9):831–838, Sep. 1986.
- S. M. Dibaji, M. Pirani, A. M. Annaswamy, K. H. Johansson, and A. Chakraborty. Secure control of wide-area power systems: Confidentiality and integrity threats. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 7269–7274, Dec 2018.
- J. C. Doyle, B. A. Francis, and A. R. Tannenbaum. *Feedback control theory*. Courier Corporation, 2013.
- F. Farokhi, I. Shames, and N. Batterham. Secure and private control using semi-homomorphic encryption. *Control Engineering Practice*, 67:13 – 20, 2017. ISSN 0967-0661.
- Z. Guo, D. Shi, K. H. Johansson, and L. Shi. Worst-case stealthy innovation-based linear attack on remote state estimation. *Automatica*, 89:117 – 124, 2018.
- P. Hespanhol, M. Porter, R. Vasudevan, and A. Aswani. Statistical watermarking for networked control systems. In *2018 Annual American Control Conference (ACC)*, pages 5467–5472, June 2018.
- C. Johnson. Accommodation of external disturbances in linear regulator and servomechanism problems. *IEEE Transactions on Automatic Control*, 16(6):635–644, December 1971.
- K. Kogiso and T. Fujita. Cyber-security enhancement of networked control systems using homomorphic encryption. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 6836–6843, Dec 2015.
- D. Kushner. The real story of stuxnet. *IEEE Spectrum*, 50(3):48–53, March 2013.
- R. M. Lee, M. J. Assante, and T. Conway. German steel mill cyber attack. *E-ISAC*, 2014.
- R. M. Lee, M. J. Assante, and T. Conway. Analysis of the cyber attack on the Ukrainian power grid. defense use case. *E-ISAC*, 2016.
- Y. Z. Lun, A. D’Innocenzo, F. Smarra, I. Malavolta, and M. D. Di Benedetto. State of the art of cyber-physical systems security: An automatic control perspective. *Journal of Systems and Software*, 149:174 – 216, 2019.
- Y. Mo and B. Sinopoli. False data injection attacks in control systems. In *First Workshop on Secure Control Systems*, April 2010.
- Y. Mo, S. Weerakkody, and B. Sinopoli. Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs. *IEEE Control Systems*, 35(1):93–109, Feb 2015.
- C. Murguia and J. Ruths. CUSUM and chi-squared attack detection of compromised sensors. In *2016 IEEE Conference on Control Applications (CCA)*, pages 474–480, Sept 2016.
- A.C.M. Ran and R. Vreugdenhil. Existence and comparison theorems for algebraic Riccati equations for continuous- and discrete-time systems. *Linear Algebra and its Applications*, 99:63 – 83, 1988. ISSN 0024-3795.
- G. Stein. Respect the unstable. *IEEE Control Systems Magazine*, 23(4):12–25, Aug 2003.
- A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson. A secure control framework for resource-limited adversaries. *Automatica*, 51:135 – 148, 2015.
- D. Umsonst and H. Sandberg. On the confidentiality of linear anomaly detector states. In *2019 American Control Conference (ACC)*, July 2019.
- M. Xue, W. Wang, and S. Roy. Security concepts for the dynamics of autonomous vehicle networks. *Automatica*, 50(3):852 – 857, 2014.
- Y. Yuan and Y. Mo. Security in cyber-physical systems: Controller design against known-plaintext attack. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 5814–5819, Dec 2015.