Preprints of the 20th World Congress
The International Federation of Automatic Control
Toulouse, France, July 9-14, 2017

Analysis and Mitigation of Bias Injection Attacks Against a Kalman Filter ⋆

**Jezdimir Milošević Takashi Tanaka Henrik Sandberg
Karl Henrik Johansson**

*ACCESS Linnaeus Center
KTH Royal Institute of Technology, 10044, Stockholm, Sweden
email: {jezdimir,ttanaka,hsan,kallej}@kth.se.*

Abstract: In this paper, we consider a state estimation problem for stochastic linear dynamical systems in the presence of bias injection attacks. A Kalman filter is used as an estimator, and a chi-squared test is used to detect anomalies. We first show that the impact of the worst-case bias injection attack in a stochastic setting can be analyzed by a deterministic quadratically constrained quadratic program, which has an analytical solution. Based on this result, we propose a criterion for selecting sensors to secure in order to mitigate the attack impact. Furthermore, we derive a condition on the necessary number of sensors to secure in order for the impact to be less than a desired threshold.

Keywords: Cyber-Physical Systems, Cyber-Security, Cyber-Attacks.

1. INTRODUCTION

Many physical processes nowadays are integrated with computing and networking devices, forming so called cyber-physical systems (CPSs). CPSs are envisioned to improve our world in many aspects. Power grids with high penetration of renewable energy, smart traffic lights that reduce congestions, and energy efficient buildings are just some of numerous examples. However, the fusion of cyber with the physical world also creates new threat. By gaining unauthorized control over some of the cyber components of a CPS, malicious attackers are able to endanger the physical world. Successful cyber-attacks could increase cost of operation, cause physical damage, and even pose a threat on a national scale (Slay and Miller, 2007; Zeller, 2011; Kushner, 2013). Therefore, in order to be able to exploit all the benefits that CPSs have to offer, the problem of cyber-security has to be addressed from all relevant fields including control engineering.

One type of cyber-attacks that attracted considerable interest within the control community are so called false-data injection attacks. In these attacks, the attacker intercepts and alters some of the measurement and/or control signals in a coordinated way. Not only can these attacks accomplish malicious goal such as destabilizing the system or considerably increase the estimation error, but they can also be designed to remain undetected by some of existing anomaly detection algorithms (Mo et al., 2010; Liu et al., 2011; Cárdenas et al., 2011; Smith, 2011; Pasqualetti et al., 2013; Amin et al., 2013; Teixeira et al., 2015; Guo et al., 2016). Various approaches for detecting these types of attacks (Teixeira et al., 2012; Pasqualetti et al., 2013; Mo et al., 2015) and mitigating their impact (Kim and Poor, 2011; Vukovic et al., 2012) have been proposed.

In one of the approaches for attack mitigation (Kim and Poor, 2011; Vukovic et al., 2012), the authors consider localizing and placing additional security measures on the most vulnerable components in the system. These studies were made in the context of state estimation of power grids. The grid was modeled as a linear static noiseless system, and a bad data detector (BDD) was used for detection of anomalies. The mitigation methods included data authentication, multi-path routing of some sensor measurements, video surveillance, and temper proof components. The authors proposed securing those sensors that make the BDD most vulnerable against undetectable bias injection attacks. Although the approach showed promising results, it is applicable only to static linear systems with BDD. The extension of this work to stochastic system models, and other types of estimators and anomaly detectors, have not been addressed so far to the best of our knowledge.

In this paper, we study the problem of security allocation in stochastic linear dynamical systems. We consider a state estimation problem where a Kalman filter is used as an estimator, and a chi-squared test is used to detect anomalies. As a first step, we focus our analysis on protection against bias injection attacks (Teixeira et al., 2015). In these attacks, the attacker's goal is to increase the mean square estimation error by adding a constant bias to some of the sensor measurements, while remaining undetected. Our aim is to find a criterion for selecting sensors to secure in order to mitigate the attack impact.

The contributions of this paper are the following. First, we extend the bias injection attack to the stochastic state estimation problem. This generalizes (Teixeira et al., 2015), where these attacks were studied in a deterministic setting, and phenomena such as false alarms in stochastic

⋆ This work was supported by the Swedish Civil Contingencies Agency through the CERCES project, the Swedish Research Council, Knut and Alice Wallenberg Foundation, and the Swedish Foundation for Strategic Research.

systems were neglected. Second, we prove in Theorem 1 that the problem of finding the worst case bias injection attack in the stochastic setting can be transformed to a quadratically constrained quadratic program, similar to the deterministic case. Based on this result, we propose a criterion for selecting sensors to secure. Third, in Theorem 2, we derive a condition on the necessary number of sensors to secure in order for the bias attack impact to be less than a desired threshold.

The paper is organized as follows. In the remainder of this section, we revisit some results from linear algebra that we use. In Section II, we introduce the problem of estimation in the presence of cyber-attacks. In Section III, we formulate the problem of finding the worst case bias injection attacks. In Section IV, we provide analysis of the attack impact, propose a criterion for selecting sensors to secure, and derive a condition on the necessary number of sensors to be secured. In Section V, we conclude the paper.

Preliminaries. We briefly revisit some results from linear algebra related to generalized eigenvalues.

Definition 1. Let M, N be matrices in $\mathbb{C}^{n\times n}$. The set of *generalized eigenvalues* of the matrix pencil(pair) (M, N) is defined as

$$\lambda(M,N) = \{\lambda \in \mathbb{C} : \det(M - \lambda N) = 0\}.$$

The *generalized eigenvector* x of (M, N) is a nontrivial solution of the equation $Mx = \lambda Nx$ with $\lambda \in \lambda(M, N)$.

We are interested in the case of real pencils, with $N \succ 0$ and $M \succeq 0$. In that case, the pencil (M, N) has exactly n real nonnegative generalized eigenvalues (Golub and Van Loan, 2012, Section 7). Moreover, the following result holds (Avron et al., 2009, Special case of Theorem 3.4).

Lemma 1. Let $M \succeq 0$, $N \succ 0$, with $M, N \in \mathbb{R}^{n\times n}$, and let $0 \le \lambda_1 \le \lambda_2 \le ... \le \lambda_n$ be the generalized eigenvalues of the pencil (M, N). Then for any $j \in \{1, ..., n\}$ we have

$$\lambda_j = \min_{\substack{U\subseteq\mathbb{R}^n \\ \dim(U)=j}} \max_{\substack{x\in U \\ x\neq 0}} \frac{x^T M x}{x^T N x},$$

where U represents a subspace of the vector space $\mathbb{R}^n$.

2. MODEL SETUP

In this section, we consider an estimation problem in the presence of cyber-attacks. The Kalman filter is used to estimate the system state, and a chi-squared test is used for detection of anomalies. At some point, the attacker starts changing the sensor measurements that it controls. Our aim is to characterize how the corrupted measurements influence the state estimate, and the signal from the anomaly detector.

2.1 Plant Model

The plant is modeled as a linear time-invariant system

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

where $x(k) \in \mathbb{R}^n$ is the system state at time step k, $v(k) \in \mathbb{R}^n$ is the process noise, $y(k) \in \mathbb{R}^m$ is the vector of sensor measurements, and $w(k) \in \mathbb{R}^m$ is the measurement noise. The processes $\{v(k)\}$ and $\{w(k)\}$ are independent, zero mean, Gaussian white processes with covariance matrices $\Sigma_v \succeq 0$ and $\Sigma_w \succ 0$, respectively. The initial state of the system $x(0)$ is a Gaussian random variable with mean value $\bar{x}(0)$ and covariance matrix $\Sigma_{x(0)} \succ 0$, independent of $\{v(k)\}$ and $\{w(k)\}$. We assume the pair (C, A) is detectable, and the pair $(\Sigma_v^{1/2}, A)$ is stabilizable.

The state vector $x(k)$ is estimated using the Kalman filter. Under stabilizability and detectability assumptions we introduced, filter reaches a steady state. The state estimate then evolves according to the equation

$$\hat{x}(k+1|k) = (A - KC)\hat{x}(k|k-1) + Ky(k), \tag{2}$$

where $\hat{x}(k|k-1)$ represents the one step ahead prediction, and

$$K = A\Sigma_e C^T (C\Sigma_e C^T + \Sigma_w)^{-1},$$

represents the steady state Kalman gain. The matrix Σ_e represents the steady state covariance matrix of the estimation error

$$e(k) = x(k) - \hat{x}(k|k-1),$$

and it is obtained by solving a steady state Riccati equation

$$\Sigma_e = A(\Sigma_e - \Sigma_e C^T (C\Sigma_e C^T + \Sigma_w)^{-1} C\Sigma_e)A^T + \Sigma_v.$$

The matrix $A - KC$ of the Kalman filter is asymptotically stable (Anderson and Moore, 2012, Chapter 4).

In order to detect possible anomalies, a chi-squared test is used. The first step of the anomaly detection procedure is to generate a residual signal

$$r(k) = y(k) - C\hat{x}(k|k-1). \tag{3}$$

Note that $C\hat{x}(k|k-1)$ is the estimate of $y(k)$, thus $r(k)$ represents the difference between $y(k)$ and its modeled behavior. In the absence of anomalies, $\{r(k)\}$ is a white Gaussian process with zero mean and covariance matrix

$$\Sigma_r = C\Sigma_e C^T + \Sigma_w.$$

The statistical approach we use assumes that the presence of anomalies would change the distribution of $r(k)$. Thus, the second step of the anomaly detection procedure is to define a suitable test to judge if the residual $r(k)$ comes from the Gaussian distribution that we mentioned previously, or if an anomaly occurred and the distribution changed. A simple method that is used for this purpose is to test if the squared distance measure

$$\chi^2(k) = r^T(k)\Sigma_r^{-1} r(k) = \|\Sigma_r^{-1/2} r(k)\|_2^2,$$

is greater than a sufficiently large threshold $\tau > 0$. The random variable $\chi^2(k)$ is distributed according to a chi-squared probability distribution, which is the reason why this test is called a chi-squared test.

In absence of anomalies, the random variable $\chi^2(k)$ takes relatively small values most of the time. The cases when this signal exceeds the threshold τ might be an indication that an anomaly occurred. However, it is important to realize that any threshold τ will occasionally be breached even when anomalies are not present. These false alarms happen due to the random noise, with probability

$$\mathbb{P}(\|\Sigma_r^{-1/2} r(k)\|_2^2 > \tau) =: \alpha.$$

Large values of τ would decrease the false alarm probability α, but also the sensitivity to anomalies. On the other hand, a small value of τ would result in high probability of false alarms, which is also undesirable. Hence, τ has

to be chosen as a reasonable trade-off between these two phenomena.

The presence of false alarms makes attack detection difficult. If an alarm occurs when the attack is present, and if the attack does not change distribution of $\chi^2(k)$ considerably, the alarm may be classified as a consequence of noise.

2.2 Attack Model

Suppose that at the time instant $k = k_0$ the attacker starts changing the values of the measurements it can influence. From that point onwards, the measurement equation becomes

$$y_a(k) = y(k) + Da(k), \quad k \geq k_0, \tag{4}$$

where the vector $a(k) \in \mathbb{R}^{m_a}$ represents the signal the attacker injects, and m_a is the number of measurements the attacker controls. The matrix $D \in \mathbb{R}^{m \times m_a}$ is a matrix that maps $a(k)$ to corresponding measurements in the following way. Denote by $j_1 < j_2 < ... < j_{m_a}$ indices of the measurements in vector $y(k)$ that are under the attacker's control. Then the elements $(j_1, 1), (j_2, 2)..., (j_{m_a}, m_a)$ of matrix D are one, and the rest are zero.

As we mentioned in the introduction, additional security measures are introduced on certain number of sensor measurements. Examples of these security measures could be encryption of the communication channel through which sensor measurements are transmitted, multi path routing of the measurements, or making the sensors harder to physically access. We assume that secured sensors cannot be corrupted by the attacker, which implies that rows of the D matrix that correspond to secured sensors are equal to zero.

In case the attack is not detected, the attacked measurements $y_a(k)$ are used to construct the state estimate. Due to the linearity of the Kalman filter, the attacked estimate $\hat{x}_a(k)$ is the sum of the responses to $y(k)$ and $a(k)$, i.e.,

$$\hat{x}_a(k) = \hat{x}(k|k-1) + \Delta\hat{x}(k), \tag{5}$$

where $\Delta\hat{x}(k)$ is dependent just on the attack signal and propagates by

$$\Delta\hat{x}(k+1) = (A - KC)\Delta\hat{x}(k) + KDa(k). \tag{6}$$

The error between state $x(k)$ and the corrupted estimate $\hat{x}_a(k)$ now becomes

$$x(k) - \hat{x}_a(k) = e(k) - \Delta\hat{x}(k). \tag{7}$$

The attack influences the residual signal $r(k)$ as well. From (3), the attacked residual signal changes to

$$y_a(k) - C\hat{x}_a(k) = r(k) + \Delta r(k),$$

where $\Delta r(k)$ can be obtained from (4) and (5) as

$$\Delta r(k) = Da(k) - C\Delta\hat{x}(k). \tag{8}$$

The goal of the attacker is to increase the mean square of estimation error (7), and at the same time remain undetected. In next section, we introduce bias injection attacks as one of the possible strategies to construct signal $a(k)$ that accomplishes this goal.

3. BIAS INJECTION ATTACKS

In the bias injection attack scenario, the attack signal $a(k)$ slowly converges to some constant vector a (Teixeira et al., 2015). In this section, we formulate the problem of finding a vector a that maximizes mean square estimation error in steady state, and does not increase alarm probability more then a certain threshold.

By substituting $a(k)$ with a in (6), we get

$$\Delta\hat{x}(k+1) = (A - KC)\Delta\hat{x}(k) + KDa.$$

As we mentioned earlier, matrix $A - KC$ of the Kalman filter is asymptotically stable, hence both $\Delta\hat{x}(k)$ and $\Delta r(k)$ reach steady states. The steady state equations for $\Delta\hat{x}(k)$ and $\Delta r(k)$ are

$$\Delta\hat{x} = (A - KC)\Delta\hat{x} + KDa, \tag{9}$$

$$\Delta r = Da - C\Delta\hat{x}. \tag{10}$$

Since $A - KC$ is stable, $I_n - A + KC$ is invertible, thus the solution of (9) is unique and given by

$$\Delta\hat{x} = (I_n - A + KC)^{-1}KDa =: G_{\hat{x}}Da. \tag{11}$$

Combining (11) with (10) gives

$$\Delta r = Da - CG_{\hat{x}}Da = (I_m - CG_{\hat{x}})Da =: G_rDa. \tag{12}$$

Remark 1. Note that the previous discussion holds in case that the bias injection attack is not detected during the transient phase. We will assume that this is not the case since the attacker can make the transient smooth by increasing the attack slowly (Teixeira et al., 2015).

We define the attacker's objective as to increase the mean square of error (7) once $\Delta\hat{x}(k)$ reaches steady state $\Delta\hat{x}$, that is

$$\underset{a \in \mathbb{R}^{m_a}}{\text{maximize}} \quad \mathbb{E}\{\|e(k) - \Delta\hat{x}\|_2^2\}. \tag{13}$$

The objective can be rewritten as

$$\mathbb{E}\{\|e(k) - \Delta\hat{x}\|_2^2\} = \mathbb{E}\{\|e(k)\|_2^2 - 2\Delta\hat{x}^T e(k) + \|\Delta\hat{x}\|_2^2\}.$$

The term $\mathbb{E}\{2e(k)^T\Delta\hat{x}\}$ is equal to zero because $\Delta\hat{x}$ is constant, and $e(k)$ is zero mean. The term $\mathbb{E}\{\|e(k)\|_2^2\} = \mathrm{Tr}(\Sigma_e)$ is constant, since it represents the part of the error coming from the noise. Thus, in order to maximize (13), the attacker needs to find a that maximizes

$$\mathbb{E}\{\|\Delta\hat{x}\|_2^2\} = \|G_{\hat{x}}Da\|_2^2.$$

The constraint for the attacker is that it wants to remain undetected. The bias injection attacks preserves the nature of residual distribution, since it remains Gaussian. However, since the attack changes the mean value of the distribution, the alarm probability can increase. For this reason, we assume that the constraint for the attacker is to not considerably increase the alarm probability. This constraint can be modeled as

$$\mathbb{P}(\|\Sigma_r^{-1/2}(r(k) + \Delta r)\|_2^2 > \tau) \leq \alpha + \Delta\alpha \leq 1,$$

where $\Delta\alpha > 0$ and is a threshold that models the attacker's willingness to risk detection. For example, say that the false alarm probability is $\alpha = 5\%$. In case that the alarm probability in presence of attacks raises to $\alpha + \Delta\alpha = 5.5\%$, the alarms will probably be classified as a consequence of noise, and the attack will remain undetected. However, in case that the alarm probability in presence of attack changes to $\alpha + \Delta\alpha = 25\%$, the attack will most likely be detected.

Based on the previous discussion, the problem the attacker wants to solve can be formalized as:

Problem 1.

$$\underset{a \in \mathbb{R}^{m_a}}{\text{maximize}} \ \|G_{\hat{x}}Da\|_2^2, \tag{14a}$$

$$\text{s.t. } \mathbb{P}(\|\Sigma_r^{-1/2}(r(k)+\Delta r)\|_2^2 > \tau) \leq \alpha + \Delta\alpha. \quad (14b)$$

Note that Problem 1 in general has many parameters that are not necessarily known to the attacker. In this paper we are interested in studying the worst case scenario, thus we adopt the following assumption about the attacker.

Assumption 1. The attacker:

- knows the structure of Problem 1;
- is able to gain control over all the sensors that are not secured.

In order to find the solution of Problem 1, we will prove that it can be transformed into

Problem 2.

$$\underset{a\in\mathbb{R}^{m_a}}{\text{maximize}} \ \|G_{\hat{x}}Da\|_2^2, \quad (15a)$$

$$\text{s.t.} \quad \|\Sigma_r^{-1/2}G_rDa\|_2^2 \leq \delta^2. \quad (15b)$$

Problem 2 is a quadratically constrained quadratic program that has an analytical solution. We will show that for a particular choice of δ, Problem 1 is equivalent to Problem 2.

The intuition behind the proof is the following. The random variable $\Sigma_r^{-1/2}r(k)$ has Gaussian distribution with *unit covariance* and zero mean. This distribution is spherically symmetric, and the alarm probability is equal to the integral of the distribution outside the spherical constraint. The bias injection attack is just affecting the mean value of the distribution, i.e., the attack shifts the distribution from zero to $\Sigma_r^{-1/2}\Delta r$. Because of the spherical symmetry of the distribution, as well as the spherical symmetry of the integrated area, the direction of the shift does not play any role. It is only the magnitude $\|\Sigma_r^{-1/2}\Delta r\|_2$ that affects the alarm rate. The following lemma is used in the proof.

Lemma 2. The alarm probability

$$\mathbb{P}(\|\Sigma_r^{-1/2}(r(k)+\Delta r)\|_2^2 > \tau),$$

is *strictly increasing* in $\|\Sigma_r^{-1/2}\Delta r\|_2^2 = \|\Sigma_r^{-1/2}G_rDa\|_2^2$.

Proof. The random variable $\Sigma_r^{-1/2}(r(k)+\Delta r)$ is Gaussian with mean value $\Sigma_r^{-1/2}\Delta r = \Sigma_r^{-1/2}G_rDa$ and covariance matrix $\mathbb{E}\{\Sigma_r^{-1/2}r(k)r^T(k)(\Sigma_r^{-1/2})^T\} = I_m$. Thus, $\|\Sigma_r^{-1/2}(r(k)+\Delta r)\|_2^2$ represents the non-central chi-squared random variable (Seber, 1963). The non-central chi-squared distribution is defined by two parameters. The first one is the number of degrees of freedom, and it is equal to the dimension of the random variable m. The second parameter is called the non-centrality parameter, and it is equal to the magnitude of the mean value $\|\Sigma_r^{-1/2}G_rDa\|_2^2$.

The alarm probability represents the complementary cumulative distribution function of the non-central chi-squared random variable. It is known that this function is equal to the Marcum Q-function (Sun et al., 2010). It was proven in the same work (Sun et al., 2010, Section III), that the generalized Marcum Q-function is strictly increasing in the non-centrality parameter for $m, \tau > 0$. In our case the non-centrality parameter equals to $\|\Sigma_r^{-1/2}G_rDa\|_2^2$, so the claim of the lemma follows. □

Using the previous result, we are ready to prove the equivalence between Problem 1 and Problem 2.

Theorem 1. If Assumption 1 holds, then there exists $\delta \in \mathbb{R}$ such that Problem 1 is equivalent to Problem 2.

Proof. Let $\bar{r} \in \mathbb{R}^m$ be any unit vector, and let δ be such that

$$\mathbb{P}(\|\Sigma_r^{-1/2}r(k) + \delta\bar{r}\|_2^2 > \tau) = \alpha + \Delta\alpha, \quad (16)$$

is satisfied. Such δ exists, since the alarm probability is the Marcum Q-function, and this function is continuous and strictly increasing in δ^2 for $m, \tau > 0$ (Sun and Baricz, 2008).

The objective functions (14a) and (15a) of the problems are the same, thus it is sufficient to prove that the feasible domains specified by the constraints are equal. We will use a contradiction argument to prove this. Assume that there exists a that satisfies constraint (14b)

$$\mathbb{P}(\|\Sigma_r^{-1/2}(r(k)+G_rDa)\|_2^2 > \tau) \leq \alpha + \Delta\alpha, \quad (17)$$

but violates constraint (15b)

$$||\Sigma_r^{-1/2}G_rDa||_2^2 > \delta^2.$$

In that case, we have that

$$||\Sigma_r^{-1/2}G_rDa||_2^2 > \delta^2 = ||\delta\bar{r}||_2^2,$$

and from (16) and (17)

$$\mathbb{P}(\|\Sigma_r^{-1/2}(r(k)+G_rDa)\|_2^2 > \tau) \leq \mathbb{P}(\|\Sigma_r^{-1/2}r(k) + \delta\bar{r}\|_2^2 > \tau).$$

This is in contradiction with Lemma 2 where it was proven that the alarm probability is strictly increasing function of $||\Sigma_r^{-1/2}G_rDa||_2^2$. In a similar manner, we disprove the existence of a that satisfies the quadratic constraint, but violates the probability constraint. □

Remark 2. Since there is no simple closed form for the Marcum Q-function, finding the δ that satisfies

$$\mathbb{P}(\|\Sigma_r^{-1/2}r(k) + \delta\bar{r}\|_2^2 > \tau) = \alpha + \Delta\alpha,$$

needs to be done either numerically, or by using a Monte Carlo method.

It is interesting to note that the optimization Problem 2 is of the same form as the one in (Teixeira et al., 2015), which was obtained by analyzing bias injection attacks in a deterministic setting.

4. ATTACK IMPACT MITIGATION

In this section, we address the problem of systematically mitigating the impact of bias injection attacks. Based on an analysis of the solution of Problem 2, we propose a criterion for selecting the best combination of sensors to secure. Furthermore, we also provide a condition on the necessary number of sensors to secure in order for the bias injection attack impact to be less than a desired threshold.

It is always of interest to first check if the attacker is able to inflict arbitrary large damage while staying undetected at the same time. The following result shows that for the problem we consider, this is possible only in a very restricted case.

Proposition 1. A necessary condition for the optimal value of Problem 2 to be unbounded is that the matrix A has an eigenvalue equal to 1.

Proof. Assume that the attacker is able to increase the error arbitrarily. A necessary condition for that is existence of $Da \neq 0$ such that $\Sigma_r^{-1/2} G_r Da = \Sigma_r^{-1/2} \Delta r = 0$. Since matrix Σ_r^{-1} is positive definite, it follows $\Delta r = G_r Da = 0$. In that case, from (10) we have $Da = C\Delta\hat{x}$. But then it follows from (9) that

$$\begin{aligned}\Delta\hat{x} &= (A - KC)\Delta\hat{x} + KDa \\ &= A\Delta\hat{x} + K(Da - C\Delta\hat{x}) = A\Delta\hat{x}.\end{aligned}$$

We see that the last equation has a nontrivial solution only in case that matrix A has an eigenvalue equal to 1. □

Corollary 1. Note that if A does not have eigenvalue equal to 1, it follows from the proof that $\text{null}(G_r) = \{\emptyset\}$.

Since matrix A has eigenvalue equal to 1 only in exceptional cases, which can be treated independently, we introduce the following assumption without significant loss of generality.

Assumption 2. We assume A does not have eigenvalue equal to 1.

Under Assumption 2, the solution of Problem 2 can be found analytically.

Lemma 3. (Teixeira et al., 2015, Theorem 11) Suppose Assumption 2 holds. The solution of Problem 2 is then given by

$$a^* = \pm \frac{\delta}{\|\Sigma_r^{-1/2} G_r D v^*\|_2} v^*,$$

where v^* is the unit length generalized eigenvector that corresponds to the maximal generalized eigenvalue λ^* of the matrix pencil

$$(D^T G_{\hat{x}}^T G_{\hat{x}} D, D^T G_r^T \Sigma_r^{-1} G_r D). \tag{18}$$

The maximal increase of the mean squared error is

$$\|G_{\hat{x}} Da^*\|_2^2 = \lambda^* \delta^2. \tag{19}$$

It follows from the result that the attack impact is the product of two essentially different parts. The first part is δ^2, which models that the attack can increase the impact by increasing the risk of being detected. This part cannot be influenced by securing sensors. The second part is λ^*, and it is dependent on the properties of the matrix pencil (18). We next explain how this matrix pencil changes by securing some of the sensors.

The only parameter that changes in the matrix pencil (18) with securing sensors is the matrix D. Assume we want to secure l sensor measurements. Under Assumption 1, matrix D has m rows and $m - l$ columns. Recall that we assumed that secured sensors cannot be corrupted by the attacker, so rows that correspond to those measurements are equal to zero. Therefore, the problem is to choose which l rows of D should be zero, such that the maximal generalized eigenvalue of the pencil (18) is minimal. This is a combinatorial problem, and can be solved by going through all possible combinations. This is feasible in practice as long as the problem size is not too large.

Besides the largest generalized eigenvalue, the other ones prove to be useful for the attack analysis. Assume that none of the sensors are secured, so that $D = I_m$ and the matrix pencil (18) is equal to $(G_{\hat{x}}^T G_{\hat{x}}, G_r^T \Sigma_r^{-1} G_r)$. Under Assumption 2, $\text{null}(G_r) = \{\emptyset\}$, which implies that $G_r^T \Sigma_r^{-1} G_r$ is positive definite and the pencil $(G_{\hat{x}}^T G_{\hat{x}}, G_r^T \Sigma_r^{-1} G_r)$ has exactly m real nonnegative generalized eigenvalues (see Preliminaries). It turns out that the impact of the worst case bias injection attack with *any* p sensors is always larger than p–th generalized eigenvalue of the matrix pencil $(G_{\hat{x}}^T G_{\hat{x}}, G_r^T \Sigma_r^{-1} G_r)$.

Theorem 2. Suppose Assumption 2 holds. Denote by $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m$ the generalized eigenvalues of $(G_{\hat{x}}^T G_{\hat{x}}, G_r^T \Sigma_r^{-1} G_r)$. Assume the attacker has control over $p \in \{1, \ldots, m\}$ sensors. Then the maximal impact (19) conducted with any combination of p sensors cannot be less than $\lambda_p \delta^2$.

Proof. In case that the attacker controls p measurements, the attack signal Da is sparse with p non-zero components. Define by $\mathbb{I}_p = \{Da \in \mathbb{R}^m | \text{card}(Da) \leq p\}$ the set of all possible bias injection attacks that the attacker is able to construct using p measurements ($\text{card}(Da)$ returns the number of non-zero elements in vector Da). Using Lemma 1, it follows

$$\lambda_p^* = \min_{\substack{U \subseteq \mathbb{I}_p \\ \dim(U) = p}} \max_{Da \in U} \frac{a^T D^T G_{\hat{x}}^T G_{\hat{x}} Da}{a^T D^T G_r^T \Sigma_r^{-1} G_r Da} \geq$$

$$\min_{\substack{U \subseteq \mathbb{R}^m \\ \dim(U) = p}} \max_{x \in U} \frac{x^T G_{\hat{x}}^T G_{\hat{x}} x}{x^T G_r^T \Sigma_r^{-1} G_r x} = \lambda_p, \tag{20}$$

since $\mathbb{I}_p \subseteq \mathbb{R}^m$. Recall that the attack impact is given by (19). By multiplying (20) with δ^2, it follows that minimal impact conducted with p sensors cannot be less than $\lambda_p \delta^2$, which completes the proof. □

In order to illustrate Theorem 2 and other results from this section, we consider the following example.

Example 1. Consider the system

$$A = \begin{bmatrix} 0.9 & 0 & 0.3 & 0 \\ 0 & 0.9 & 0 & 0.3 \\ 0 & 0 & 0.9 & 0 \\ 0 & 0 & 0 & 0.9 \end{bmatrix} \quad \Sigma_v = 10^{-2} \begin{bmatrix} 0.8 & 0.2 & 2.7 & 0.7 \\ 0.2 & 0.8 & 0.7 & 2.7 \\ 2.7 & 0.7 & 9.0 & 2.3 \\ 0.7 & 2.7 & 2.3 & 9.0 \end{bmatrix}$$

$$C = I_4 \qquad \Sigma_w = \text{diag}(2.5, 2.5, 0.5, 0.5).$$

We assume for simplicity that $\delta^2 = 1$, so the attack impact (19) is equal to the largest generalized eigenvalue.

Assume first that none of the sensors are secured, that is, $D = I_m$ and the pencil (18) is equal to $(G_{\hat{x}}^T G_{\hat{x}}, G_r^T \Sigma_r^{-1} G_r)$. The generalized eigenvalues are then $\lambda_1 = 0.02$, $\lambda_2 = 0.02$, $\lambda_3 = 66.05$, and $\lambda_4 = 115.72$. Thus, in case we do not secure any sensors, the worst case impact is 115.72.

Assume now we want to secure a certain number of sensors, so that impact is less than, say 5 for $\delta^2 = 1$. We can then use Theorem 2 to decide the necessary number of sensors to secure. Since the second highest generalized eigenvalue is $\lambda_3 = 66.05$, it follows from Theorem 2 that we cannot reduce the impact to be less than that value if we secure only one sensor. We also see that the third largest eigenvalue is equal to 0.02, so securing two senors could be more beneficial.

The maximal generalized eigenvalues for different combination of attacked measurements are shown in Table 1. We can see from the table that by securing sensors $\{3, 4\}$ we achieve the best result, reducing the maximal impact

Table 1. Maximal Generalized Eigenvalue for Different Combinations of the Secured Sensors

Secured sensors	λ^*	Secured sensors	λ^*
$\emptyset$	115.72	{2,3}	2.15
{1}	84.86	{2,4}	84.79
{2}	84.86	{3,4}	1.45
{3}	85.01	{1,2,3}	2.15
{4}	85.01	{1,2,4}	2.15
{1,2}	2.27	{1,3,4}	1.43
{1,3}	84.79	{2,3,4}	1.43
{1,4}	2.15	{1,2,3,4}	0

by $115.72/1.45 \approx 80$ times. Note as well that the bound from Theorem 2 could be quite loose. For example, the maximal attack conducted with two sensors is 84.79, while the bound from Theorem 2 is equal to $\lambda_2 = 0.02$. □

5. CONCLUSIONS AND FUTURE WORK

In this paper, we considered the problem of bias injection attacks against the Kalman filter equipped with the chi-squared detector. We proved that the problem of finding a worst-case bias injection attack can be reduced to quadratically constrained quadratic program. Based on the analysis of the solution, we derived a criterion based on which we select sensors to secure. We also derived a condition on the necessary number of sensors to secure in order for the bias attack impact to be less than a desired threshold.

There are two main challenges that we plan to address in the future work. Firstly, we plan to extend the problem of protection to more general types of attacks. Secondly, we saw that the problem of securing sensors was combinatorial in nature. For large-scale systems, it may be computationally expensive to check all possibilities, and algorithms that could solve these particular combinatorial problems fast might be required.

REFERENCES

Amin, S., Litrico, X., Sastry, S., and Bayen, A.M. (2013). Cyber security of water SCADA systems – Part I: Analysis and experimentation of stealthy deception attacks. *IEEE Transactions on Control Systems Technology*, 21(5), 1963–1970.

Anderson, B.D. and Moore, J.B. (2012). *Optimal filtering*. Courier Corporation.

Avron, H., Ng, E., and Toledo, S. (2009). Using perturbed QR factorizations to solve linear least-squares problems. *SIAM Journal on Matrix Analysis and Applications*, 31(2), 674–693.

Cárdenas, A.A., Amin, S., Lin, Z.S., Huang, Y.L., Huang, C.Y., and Sastry, S. (2011). Attacks against process control systems: Risk assessment, detection, and response. In *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, 355–366. ACM.

Golub, G.H. and Van Loan, C.F. (2012). *Matrix computations*, volume 3. JHU Press.

Guo, Z., Shi, D., Johansson, K.H., and Shi, L. (2016). Optimal linear cyber-attack on remote state estimation. *IEEE Transactions on Control of Network Systems*, PP(99), 1–1.

Kim, T.T. and Poor, H.V. (2011). Strategic protection against data injection attacks on power grids. *IEEE Transactions on Smart Grid*, 2(2), 326–333.

Kushner, D. (2013). The real story of Stuxnet. *IEEE Spectrum*, 3(50), 48–53.

Liu, Y., Ning, P., and Reiter, M.K. (2011). False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security (TISSEC)*, 14(1), 13.

Mo, Y., Garone, E., Casavola, A., and Sinopoli, B. (2010). False data injection attacks against state estimation in wireless sensor networks. In *49th IEEE Conference on Decision and Control (CDC)*, 5967–5972.

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

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

Seber, G. (1963). The non-central chi-squared and beta distributions. *Biometrika*, 50(3/4), 542–544.

Slay, J. and Miller, M. (2007). Lessons learned from the Maroochy water breach. In *International Conference on Critical Infrastructure Protection*, 73–82. Springer.

Smith, R.S. (2011). A decoupled feedback structure for covertly appropriating networked control systems. *IFAC Proceedings Volumes*, 44(1), 90–95.

Sun, Y., Baricz, Á., and Zhou, S. (2010). On the monotonicity, log-concavity, and tight bounds of the generalized Marcum and Nuttall Q-functions. *IEEE Transactions on Information Theory*, 56(3), 1166–1186.

Sun, Y. and Baricz, Á. (2008). Inequalities for the generalized Marcum Q-function. *Applied Mathematics and Computation*, 203(1), 134 – 141.

Teixeira, A., Shames, I., Sandberg, H., and Johansson, K.H. (2012). Revealing stealthy attacks in control systems. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, 1806–1813. IEEE.

Teixeira, A., Shames, I., Sandberg, H., and Johansson, K.H. (2015). A secure control framework for resource-limited adversaries. *Automatica*, 51, 135–148.

Vukovic, O., Sou, K.C., Dan, G., and Sandberg, H. (2012). Network-aware mitigation of data integrity attacks on power system state estimation. *IEEE Journal on Selected Areas in Communications*, 30(6), 1108–1118.

Zeller, M. (2011). Myth or reality - does the Aurora vulnerability pose a risk to my generator? In *Protective Relay Engineers, 2011 64th Annual Conference for*, 130–136.