Security for control systems under sensor and actuator attacks

Hamza Fawzi, Paulo Tabuada, Suhas Diggavi

Department of Electrical Engineering,
University of California at Los Angeles, Los Angeles, CA 90095-1594
{hfawzi,tabuada,suhas}@ee.ucla.edu

***Abstract*— We consider the problem of estimation and control of a linear system when some of the sensors or actuators are attacked by a malicious agent. In our previous work [1] we studied systems with no control inputs and we formulated the estimation problem as a dynamic error correction problem with sparse attack vectors. In this paper we extend our study and look at the role of inputs and control. We first show that it is possible to increase the resilience of the system to attacks by changing the dynamics of the system using state-feedback while having (almost) total freedom in placing the new poles of the system. We then look at the problem of stabilizing a plant using output-feedback despite attacks on sensors, and we show that a principle of separation of estimation and control holds. Finally we look at the effect of attacks on actuators in addition to attacks on sensors: we characterize the resilience of the system with respect to actuator and sensor attacks and we formulate an efficient optimization-based decoder to estimate the state of the system despite attacks on actuators and sensors.**

I. Introduction

Modern control systems pervade the critical infrastructure we all depend on and have become accustomed to. Oil refineries, water distribution networks, gas networks, the power grid are just some examples of many processes in which control plays a crucial role. These systems are now becoming increasingly decentralized and geographically distributed, and thus rely on a communication network that is sometimes connected to the corporate intranet or even to the internet. This modern architecture leads to increased vulnerabilities of the system to exterior attacks that can cause physical damage to the system.

As a consequence there has been a recent surge of interest in security for cyber-physical systems [2], [3], [4], [5], [6], [7], [8]. Different approaches have been followed in order to deal with the problem. In one line of work, a static model of the plant is assumed to be known and is used to detect and identify attacks. For example in [4] the authors rely on a probabilistic model of the plant constructed from historical data and propose a novel statistical test to detect and identify stealthy attacks. In [3] the authors consider the more specific problem of attacks on power networks and give conditions under which an attack will be undetected by the *Power System State Estimator* which is part of the SCADA system that operates the network. The *Power System State Estimator* is a least-squares estimator that uses measurements from sensors and a known static model of the power network.

In another line of work, some authors have exploited the dynamics of the plant in order to deal with attacks. In [5] the authors study the problem of identifiability and detectability of attacks on sensors and actuators for systems that evolve according to a linear descriptor model. The authors characterize the fundamental limitations of any attack detection filter and show that dynamic filters are more powerful than their static counterparts. In [6], a game-theoretic approach was considered where the attacker can jam the communication link between the controller and the actuator for a limited number of time slots. The authors showed the existence of saddle-point equilibrium for this dynamic zero-sum game and derived the optimal jamming strategy for a particular instance of the problem. Finally, in [7], the problem of detecting malicious behavior is studied within the framework of the *Wireless Control Network* (see [9]). The authors characterize the nodes of the network that need to be monitored by any Intrusion Detection System in order to identify any malicious behavior in the network as well as to reconstruct the outputs of the plant (for data-logging purposes).

In our previous work [1] we adopted a novel point of view inspired from error-correction over the reals [10] which allowed us to propose a new estimation algorithm that is robust against the attacks and that is also computationally efficient, unlike most of the filters proposed previously. In our formulation, the attacks were modeled as sparse vectors that affect the output equation with the sparsity pattern being the attacked/unattacked sensors. We characterized the resilience of a system in terms of the number of sensors that can be attacked without compromising the ability to estimate the state from the corrupted measurements. Then based on techniques used in compressed sensing and error correction over the reals [10], we proposed a computationally tractable decoding algorithm that recovers the state of the system despite attacks on sensors.

In this paper we extend our study of the problem and look at the role of control and inputs. We first show that if one is allowed to implement a state-feedback law (i.e., replace the matrix A of the system by $A + BK$ for some K), then one can make the system more resilient to sensor attacks by an appropriate choice of K; furthermore one has (almost) total freedom in choosing the eigenvalues of $A + BK$ like in the

This work was partially sponsored by the NSF award 1136174 and DARPA under agreement number FA8750-12-2-0247. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA or the U.S. Government.

well-known pole placement result[1].

We then look at a problem of separation of estimation and control: we show that if there exists an output-feedback law that can stabilize the plant (with a fast enough rate) despite any attacks on q sensors, then there exists an estimation algorithm that can estimate the state of the plant despite any attacks on q sensors. In other words, when designing an output-feedback law, there is no loss of resilience in looking for a feedback law that is the composition of a state estimator with a standard state-feedback law.

Finally we look at the problem of attacks on the actuators (in addition to attacks on the sensors) and we characterize the resilience of the system to such attacks. We also propose an optimization-based decoder for such attacks.

The paper is organized as follows. We start in section II by briefly reviewing the problem formulation that we used in our previous work [1]. In section III we look at the question of increasing the resilience of a system by implementing a state feedback law. Section IV then deals with the problem of separation of estimation and control. Finally, in section V we deal with the problem of attacks on actuators.

II. Problem formulation

In this section we briefly review the problem formulation that we used in our previous work [1]. Consider a plant with state $x^{(t)} \in \mathbb{R}^n$ and sensor measurements $y^{(t)} \in \mathbb{R}^p$ evolving according to:

$$\begin{aligned} x^{(t+1)} &= Ax^{(t)} \\ y^{(t)} &= Cx^{(t)} + e^{(t)} \end{aligned} \tag{1}$$

where A is an $n \times n$ matrix, C a $p \times n$ matrix, and $e^{(t)} \in \mathbb{R}^p$ is the vector of attacks: if sensor $i \in \{1, \dots, p\}$ is not attacked then the i'th component of the vector $e^{(t)}$ is zero; otherwise sensor i is attacked and $e_i^{(t)}$ can be arbitrary. Therefore if $K \subseteq \{1, \dots, p\}$ is the set of attacked sensors, we have $\text{supp}(e^{(t)}) \subseteq K$ where $\text{supp}(e^{(t)}) = \{i \in \{1, \dots, p\} \mid e_i^{(t)} \neq 0\}$ is the support of $e^{(t)}$, i.e., the set of nonzero components of $e^{(t)}$.

We say that q errors are correctable if one can identify the state of the system from the observations even if q sensors are attacked. More formally, we have:

Definition 1: We say that q errors are correctable after T steps if there exists a decoder $D : (\mathbb{R}^p)^T \to \mathbb{R}^n$ such that for any $x^{(0)} \in \mathbb{R}^n$ and for any attack sequence $e^{(0)}, \dots, e^{(T-1)}$ with $\text{supp}(e^{(t)}) \subseteq K$ with $|K| \leq q$, we have $D(y^{(0)}, \dots, y^{(T-1)}) = x^{(0)}$ where $y^{(t)} = CA^t x^{(0)} + e^{(t)}$.

The definition above can be related to the notion of strong observability for linear systems. The important difference though is that strong observability requires the decoder to be able to recover the state for *any* attack vectors $e^{(0)}, \dots, e^{(T-1)} \in \mathbb{R}^p$, while Definition 1 only concerns attack vectors $e^{(0)}, \dots, e^{(T-1)}$ that are q-sparse. Note that strong observability clearly does not hold for the system (1) since the output disturbance matrix of the system (usually denoted by "D" in state-space representations) is equal to the identity matrix.

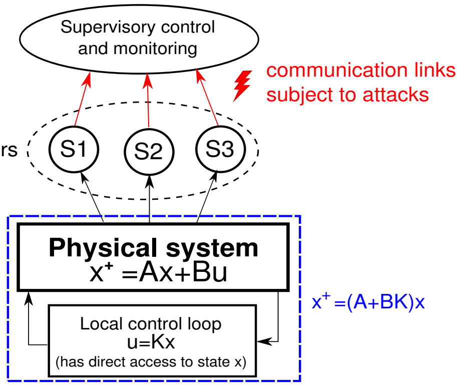
Supervisory control and monitoring
communication links subject to attacks
Sensors
S1
S2
S3
Physical system
x⁺ =Ax+Bu
x⁺ =(A+BK)x
Local control loop
u=Kx
(has direct access to state x)

Fig. 1. Scenario where a local control loop has direct access to uncorrupted sensor information. Using this local control loop, the evolution of the physical system will be governed by the matrix $A + BK$ where K can be chosen arbitrarily. The objective is to find K such that the pair $(A+BK, C)$ is resilient against a large number of attacks, which will allow the higher level supervisory system to recover the correct state despite attacks in the communication links between the sensors and the supervisory system.

The largest q such that q errors are correctable is, by definition, the resilience of a system to attacks. The following proposition from [1] gives necessary and sufficient condition on the pair (A, C) for q errors to be correctable.

Proposition 1: (from [1, Proposition 2]) For a system (A, C) and $T > 0$, the following are equivalent:

(i) q errors are correctable after T steps

(ii) For any $x \in \mathbb{R}^n \backslash \{0\}$, we have $|\text{supp}(Cx) \cup \cdots \cup \text{supp}(CA^{T-1}x)| > 2q$.

Observe from the above proposition that the largest q such that q errors are correctable is necessarily less than $\lceil p/2 - 1 \rceil$.

III. Increasing the resilience by state-feedback

In this section we look at how one can increase the resilience of a system (A, C) if one has control over its dynamics. More specifically, if B is some given matrix, we look at the problem of designing a matrix K so that the pair $(A + BK, C)$ is resilient against a large number of attacks, while it satisfies at the same time other design constraints.

From a practical point of view, this question can be motivated by the following scenario depicted in figure 1: consider a physical system that possesses a local control loop which has direct access to the state of the plant and can control its evolution. This is possible for example if the sensors are connected to the local controller through a secured wired link that is not subject to external attacks. If the local control loop implements a feedback law of the form $u = Kx$ then the evolution of the physical system is governed by the matrix $A + BK$. Also a high-level supervisory and monitoring system receives measurements from the sensors through wireless and vulnerable communication links that are subject to attacks (cf. figure 1). Observe that the choice K of the local controller will affect the resilience of the system to attacks, i.e., how many errors are correctable by the supervisory system. The objective here is therefore to

[1] The assumption that one can implement a state-feedback law might seem far-fetched since the original goal is precisely to estimate the state of the system despite the attacks. We refer the reader to section III and figure 1 for a more thorough discussion and for an example where this might be relevant.

design K in order to make the number of correctable errors of the pair $(A+BK, C)$ as large as possible.

Note that there are other design constraints that come into play in the choice of the local feedback law. Typically K is chosen so that the eigenvalues of $A+BK$ are inside the unit disc and the resulting closed-loop system is stable. It is known by the pole placement theorem that this is possible if the pair (A, B) is controllable.

In this section we ask if one can also enforce the number of correctable errors of the new pair $(A+BK, C)$ to be large, without losing the freedom of choosing the eigenvalues of $A+BK$. We show that the answer is yes, and that if the pair (A, B) is controllable, then it is possible to choose K such that $\lceil p/2-1 \rceil$ errors are correctable for $(A+BK, C)$ and such that the eigenvalues of $A+BK$ are in any arbitrary (or almost arbitrary) prescribed locations in the complex plane. In other words, by an adequate choice of the local control law, one can make the system *more resilient to attacks* (the number of correctable errors $\lceil p/2-1 \rceil$ is the maximum possible), without compromising the performance of the control.

More specifically, we show the following result:

Proposition 2: Let $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times 1}$ and $C \in \mathbb{R}^{p\times n}$ and assume that the pair (A, B) is controllable. Then for any choice of n numbers $\lambda_1, \ldots, \lambda_n \in \mathbb{C}\backslash F$ (where F is some finite subset of $\mathbb{C}$) such that the λ_i's have distinct magnitudes, there exists $K \in \mathbb{R}^{1\times n}$ such that:

- the eigenvalues of the closed-loop matrix $A+BK$ are $\lambda_1, \ldots, \lambda_n$.
- the number of correctable errors after n steps for the pair $(A+BK, C)$ is maximal (equal to $\lceil p/2-1 \rceil$).

In order to prove this result, we will make use of this lemma:

Lemma 1: Let $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{p\times n}$. Assume that A is diagonalizable and that the eigenvalues of A have distinct magnitudes. Then the following are equivalent:
(i) q errors are correctable for (A, C) after n steps.
(ii) For every eigenvector v of A, $|\mathsf{supp}(Cv)| > 2q$.

Proof:

- (i) $\Rightarrow$ (ii): This direction simply corresponds to taking x to be an eigenvector of A in the condition $|\mathsf{supp}(Cx) \cup \cdots \cup \mathsf{supp}(CA^{n-1}x)| > 2q$ of Proposition 1.
- (ii) $\Rightarrow$ (i): We assume that all eigenvectors v of A satisfy $|\mathsf{supp}(Cv)| > 2q$ and we will show that for any $x \neq 0$, we have $|\mathsf{supp}(Cx) \cup \mathsf{supp}(CAx) \cup \ldots \mathsf{supp}(CA^{n-1}x)| > 2q$. Let $x \neq 0$. Since A is diagonalizable, we can write x as a linear combination of eigenvectors of A: $x = \sum_{i=1}^{s} \alpha_i v_i$ with $\alpha_i \neq 0$ and where $v_1, \ldots, v_s$ are eigenvectors of A associated with eigenvalues $\lambda_1, \ldots, \lambda_s$. Since the eigenvalues of A have distinct magnitudes we can assume that $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_s|$. We will isolate the largest eigenvalue in this decomposition and we will call $\lambda = \lambda_1$ and $w = \alpha_1 v_1$. Now we have $(A^t x - \lambda^t w)/\lambda^t \to 0$ when $t \to +\infty$. Let $S = \mathsf{supp}(Cw)$. Note that since w is an eigenvector of A we have $|S| > 2q$ (by assumption). We'll now show that for t large enough, the support of $CA^t x$ contains S: let $\beta = \min_{i\in S} |(Cw)_i|$ and observe that clearly $\beta > 0$. Let t be large enough so that $\frac{|C(A^t x - \lambda^t w)|_i}{|\lambda|^t} < \beta/2$ for all $i \in S$. Now we have, for $i \in S$:
$$\begin{aligned}\frac{1}{|\lambda|^t}|CA^t x|_i &\geq \frac{1}{|\lambda|^t}(|\lambda^t Cw|_i - |CA^t x - \lambda^t Cw|_i)\\ &> \beta - \beta/2 = \beta/2 > 0.\end{aligned}$$
Hence $S \subset \mathsf{supp}(CA^t x)$. Thus, since $|S| > 2q$, q errors are correctable after t steps, for t large enough. But we know that the number of correctable errors does not change after n steps (by the Cayley Hamilton theorem), and so q errors are necessarily correctable after n steps. This finishes the proof.

■

We will now use this lemma to prove Proposition 2:

Proof: (Proof of Proposition 2)
To prove the result, we will show that if the chosen poles $\lambda_1, \ldots, \lambda_n$ have distinct magnitudes and do not fall in some finite set F, then there is a choice of $K \in \mathbb{R}^{1\times n}$ such that the eigenvalues of $A+BK$ are exactly the $\lambda_1, \ldots, \lambda_n$, and the corresponding eigenvectors v_i are such that $|\mathsf{supp}(Cv_i)| = p$. Thus, by the previous lemma, this will show that the number of correctable errors for $(A+BK, C)$ is $\lceil p/2-1 \rceil$.

First note that if λ is an eigenvalue of $A+BK$ and x is a corresponding eigenvector, then we have $Ax + BKx = \lambda x$, or, if $(\lambda I - A)^{-1}$ is well defined, $x = (\lambda I - A)^{-1}BKx$, i.e., x is proportional to the vector $(\lambda I - A)^{-1}B$. This means that if λ is an eigenvalue of $A+BK$, then necessarily the corresponding eigenvector is $(\lambda I - A)^{-1}B$.

We will therefore look for values of λ for which $C(\lambda I - A)^{-1}B$ has full support.

Let $i \in \{1, \ldots, p\}$ be fixed and denote by e_i the vector in $\mathbb{R}^p$ whose ith component is equal to 1 and zeros otherwise. Note that since (A, B) is controllable there exists λ such that $e_i^T C(\lambda I - A)^{-1}B \neq 0$ (see [11, Chapter 3, Theorem 2.17(ii)]), and in fact the set $F_i = \{\lambda \mid e_i^T C(\lambda I - A)^{-1}B = 0\} \subseteq \mathbb{C}$ is finite (zeros of a non-identically-zero rational fraction).

Now consider $F = (\cup_{i=1}^{n} F_i)$, and let $\lambda_1, \ldots, \lambda_n$ be any choice of n numbers in $\mathbb{C}\backslash F$ with distinct magnitudes. We will show that there exists K such that the eigenvalues of $A+BK$ are the $\lambda_j's$ and the eigenvectors v_j are such that Cv_j has full support.

By controllability of (A, B) there is a $K \in \mathbb{R}^{1\times n}$ such that the eigenvalues of $A+BK$ are the λ_j's. We know that the eigenvectors of $A+BK$ are the $v_j = (\lambda_j I - A)^{-1}B$. Now by the choice of the λ_j's and by the definition of F we know that for all j and for any i, $e_i^T C(\lambda_j I - A)^{-1}B \neq 0$. In other words, for any j, the vector Cv_j has full support. Hence, by lemma 1, the number of correctable errors of $(A+BK, C)$ is maximal. ■

IV. Separation of estimation and control

Consider a linear control system with output feedback of the form:
$$\begin{aligned}x^{(t+1)} &= Ax^{(t)} + BU^{(t)}(y^{(0)}, \ldots, y^{(t)})\\ y^{(t)} &= Cx^{(t)} + e^{(t)}\end{aligned} \qquad (2)$$

where $(U^{(t)})_{t=0,1,\ldots}$ is a (possibly dynamic) output-feedback law and $e^{(t)}$ are the attack vectors as before.

One of the main questions that we address in this section is to determine whether for a given triple (A, B, C), there exists a control law (i.e., a family $(U^{(t)})_{t=0,1,\ldots}$) that drives the state of the system (2) to the origin even if some of the sensors are attacked (i.e., that *stabilizes* the system despite attacks on some of the sensors). Observe that an attack on the sensors will affect the value of the control inputs (since the control inputs are function of the $y^{(t)}$s) which can in turn deviate the state $x^{(t)}$ from its nominal path.

It is clear that if q errors are correctable (in the sense of Definition 1, i.e., that it is possible to recover the state despite any attacks on q sensors), then one can stabilize the system in the presence of attacks. Indeed, one can simply decode the state (since q errors are correctable), and then apply a standard *state* feedback law of the form $u = Kx$ (for example). The main result of this section is to show that the converse of this statement is essentially true. More specifically, we show in Theorem 1 that if $(U^{(t)})_{t=0,1,\ldots}$ is any feedback law that stabilizes the system (with a fast enough decay) despite attacks on any q sensors, then necessarily q errors are correctable. This theorem shows that one can essentially decouple the problem of estimation and of control: there is no loss of resilience in searching for an output feedback law that is the composition of a state estimator with a standard *state* feedback.

A. Some notations and preliminaries

Before stating the separation result we will first define the notion of correctability of q errors for systems with output-feedback control inputs. We will use the symbol $E_{q,T}$ to denote the set of attack sequences of length T on any q sensors:

$$E_{q,T} = \Big\{(e^{(0)}, \ldots, e^{(T-1)}) \in (\mathbb{R}^p)^T \mid \\ \forall t \in \{0, \ldots, T-1\},\ \mathsf{supp}(e^{(t)}) \subset K \text{ and } |K| = q\Big\}.$$

We also use the notation $y(t, x^{(0)}, e)$ to denote the output at time t of the control system (2) when the initial state is $x^{(0)}$ and for the attack sequence $e \in E_{q,T}$. We now give the definition of correctability of q errors for systems with output-feedback control inputs:

Definition 2: Let a control system of the form (2) be given (this corresponds to the given of A, B, C and the control law $(U^{(t)})_{t=0,1,\ldots}$). We say that q errors are correctable after T steps if there exists a function $D : (\mathbb{R}^p)^T \to \mathbb{R}^n$ such that for any $x^{(0)} \in \mathbb{R}^n$ and any attack sequence $e \in E_{q,T}$, we have $D\Big(y(0, x^{(0)}, e), \ldots, y(T-1, x^{(0)}, e)\Big) = x^{(0)}$.

It is not hard to see that, since the systems we consider are linear and since the control inputs only depend on the measurements, the property of correctability of q errors just defined above does not depend on the control law or on B, and in fact only depends on A and C. Indeed, saying that q errors are not correctable (after T steps) for the controlled system $(A, B, C, (U^{(t)})_{t=0,1,\ldots})$ means there exists $x_a \neq x_b$, and error vectors $e_a, e_b \in E_{q,T}$ such that $y(t, x_a, e_a) = y(t, x_b, e_b)$ for all $t = 0, \ldots, T-1$. In other words, we have, for all $t \in \{0, \ldots, T-1\}$:

$$\begin{aligned} & CA^t x_a + C[B, AB, \ldots, A^{t-1}B] \begin{bmatrix} u_a^{(t-1)} \\ \vdots \\ u_a^{(0)} \end{bmatrix} + e_a^{(t)} \\ &= CA^t x_b + C[B, AB, \ldots, A^{t-1}B] \begin{bmatrix} u_b^{(t-1)} \\ \vdots \\ u_b^{(0)} \end{bmatrix} + e_b^{(t)} \end{aligned} \tag{3}$$

where

$$u_a^{(\tau)} = U^{(\tau)}(y(0, x_a, e_a), \ldots, y(\tau, x_a, e_a))$$

and

$$u_b^{(\tau)} = U^{(\tau)}(y(0, x_b, e_b), \ldots, y(\tau, x_b, e_b))$$

for $\tau = 0, \ldots, t-1$. Now observe that the terms on the left-hand side and right-hand side of (3) with the control inputs are equal (since $y(s, x_a, e_a) = y(s, x_b, e_b)$ for all s and thus $u_a^{(\tau)} = u_b^{(\tau)}$). Hence the equality (3) is equivalent to saying that for all $t \in \{0, \ldots, T-1\}$, we have:

$$CA^t x_a + e_a^{(t)} = CA^t x_b + e_b^{(t)}.$$

And this exactly means that q errors are not correctable for (A, C). This therefore shows that the notion of correctability does not depend on the control law used.

In other words, one can use the conditions developed earlier for correctability of q errors for linear systems with no inputs and apply them to systems with output-feedback control inputs. For example we have that q errors are correctable for the control system (2) if, and only if, $|\mathsf{supp}(Cx) \cup \cdots \cup \mathsf{supp}(CA^{T-1}x)| > 2q$ for all $x \neq 0$.

B. Separation of estimation and control

We are now ready to prove our result on separation of estimation and control.

Theorem 1: Let A, B, C be three matrices of appropriate sizes and assume that a control strategy given by the $(U^{(t)})_{t=0,1,\ldots}$ is such that: *for any* $x^{(0)} \in \mathbb{R}^n$ and *for any* sequence of error vectors $e \in E_{q,T}$, the sequence $(x^{(t)})$ defined by:

$$\begin{aligned} x^{(t+1)} &= Ax^{(t)} + BU^{(t)}(y^{(0)}, \ldots, y^{(t)}) \\ y^{(t)} &= Cx^{(t)} + e^{(t)} \end{aligned} \tag{4}$$

satisfies

$$\|x^{(t)}\| \leq \kappa\alpha^t \|x^{(0)}\|$$

where $\kappa > 0$ and where $0 \leq \alpha < 1$ is small enough: $\alpha < \min\{|\lambda| \mid \lambda \text{ eigenvalue of } A\}$. *Then* necessarily q errors are correctable after n steps.

Proof: We proceed by contradiction. Assume that q errors are not correctable after n steps. Then this means there exists a nonzero initial state $\bar{x} \neq 0$ that is confusable with the initial state 0. In other words, there exist attack sequences $(e_a^{(t)})_{t=0,1,\ldots}$ and $(e_b^{(t)})_{t=0,1,\ldots}$ on q sensors such that the outputs of the control system (4) in the two different executions:

1) $x^{(0)} = \bar{x}$ and $e^{(t)} = e_a^{(t)}$; and
2) $x^{(0)} = 0$ and $e^{(t)} = e_b^{(t)}$.

are equal for all $t = 0, 1, \ldots$ [2]. Now since the control law $(U^{(t)})_{t=0,1,\ldots}$ only depends on the outputs, this means that in these two executions, the same sequence of inputs, $u^{(t)}$, will be used.

Furthermore, since we must have in both cases, $\|x^{(t)}\| \leq \kappa e^{-\alpha t}\|x^{(0)}\|$, this leads, for the case where $x^{(0)} = 0$, that $x^{(t)} = 0$ for all $t \geq 0$, and so necessarily, $Bu^{(t)} = x^{(t+1)} - Ax^{(t)} = 0$ for all $t \geq 0$. Hence for the first case (when $x^{(0)} = \bar{x}$), the recurrence relation is $x^{(t+1)} = Ax^{(t)}$, which gives $x^{(t)} = A^t\bar{x}$. We now get a contradiction since $x^{(t)}$ should decay at rate of α, but the eigenvalues of A are all strictly larger than α. This completes the proof. ■

Remark: Note that the assumption on the decay rate to be fast enough is necessary; otherwise the result is not true. Indeed, if for example A is already a stable matrix, one cannot deduce anything from the mere existence of a stabilizing control law (since the system is by itself stable!). For a concrete example, take $A = 0.5I$, $B = I$, $C = I$ (note that A is stable). We know from the characterization of the number of correctable errors that even one error is not correctable after any number of steps (for example if we take $x = (1, 0, \ldots, 0)$, then $|\mathsf{supp}(Cx) \cup \mathsf{supp}(CAx) \cup \ldots| = 1 \not> 2q$ if $q > 0$). Now if we consider the trivial output feedback law $U^{(t)} = 0$ for all t, the resulting system is of course stable despite any number of attacks (the state evolution is simply $x^{(t+1)} = 0.5x^{(t)}$ and does not even depend on the sensor outputs), but as we just saw one cannot even construct a decoder to correct even one error!

V. Attacks on actuators

In this section we incorporate into our model attacks on actuators (in addition to attacks on sensors) and we study the resilience of linear control systems to such attacks. Consider a plant that evolves according to the equations:

$$\begin{aligned} x^{(t+1)} &= Ax^{(t)} + B(U^{(t)}(y^{(0)}, \ldots, y^{(t)}) + w^{(t)}) \\ y^{(t)} &= Cx^{(t)} + e^{(t)} \end{aligned} \tag{5}$$

where $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{p\times n}$ and $(U^{(t)})_{t=0,1,\ldots}$ is an output-feedback control law. As before the vectors $e^{(t)}$ represent attacks on sensors. The vectors $w^{(t)}$ represent attacks on actuators: if actuator $i \in \{1, \ldots, m\}$ is not attacked, then $w_i^{(t)} = 0$, otherwise actuator i is attacked and $w_i^{(t)}$ can be arbitrary. The set of attacked actuators will typically be denoted by L. We will also use the letter q to denote the total number of attacked nodes (sensors and actuators), $q = |K| + |L|$.

Our objective is to monitor the state of the plant from the observations $y^{(t)}$. More formally if T is some time horizon, we wish to reconstruct the sequence of states $x^{(0)}, \ldots, x^{(T-1)}$ from the observations $y^{(0)}, \ldots, y^{(T-1)}$. Observe that reconstructing the sequence $x^{(0)}, \ldots, x^{(T-1)}$ is equivalent to reconstructing the initial condition $x^{(0)}$ and the vectors $Bw^{(0)}, \ldots, Bw^{(T-2)}$. This reconstruction is possible if, and only if, the map that sends the tuple $(x^{(0)}, Bw^{(0)}, \ldots, Bw^{(T-2)}, e^{(0)}, \ldots, e^{(T-1)})$ to the corresponding outputs $(y^{(0)}, \ldots, y^{(T-1)})$ is injective [3]. Using the notations

$$\mathcal{O}_T = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{T-1} \end{bmatrix}, \quad \mathcal{M}_T = \begin{bmatrix} 0 & \ldots & \ldots & 0 \\ C & 0 & \ldots & 0 \\ CA & C & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ CA^{T-2} & CA^{T-3} & \ldots & C \end{bmatrix}$$

this map is given by:

$$\begin{bmatrix} x^{(0)} \\ (Bw^{(t)}) \\ (e^{(t)}) \end{bmatrix} \mapsto \mathcal{O}_T x^{(0)} + \mathcal{M}_T \begin{bmatrix} Bw^{(0)} \\ \vdots \\ Bw^{(T-2)} \end{bmatrix} + \begin{bmatrix} e^{(0)} \\ \vdots \\ e^{(T-1)} \end{bmatrix}$$

If this map is injective when the $w^{(t)}$'s and $e^{(t)}$'s are restricted to have less than q nonzero components combined (i.e., $|K| + |L| \leq q$), we say that q attacks are correctable, or that the system is resilient against q attacks. More formally we have the following definition:

Definition 3: Let a control system of the form (5) be given. We say that the system is resilient against q attacks after T steps if there exists a decoder $D : (\mathbb{R}^p)^T \to (\mathbb{R}^n)^T$ such that for any $x^{(0)} \in \mathbb{R}^n$, for any $w^{(0)}, \ldots, w^{(T-2)}$ with $\mathsf{supp}(w^{(t)}) \subseteq L$ and any $e^{(0)}, \ldots, e^{(T-1)}$ with $\mathsf{supp}(e^{(t)}) \subseteq K$ with $|K| + |L| \leq q$ we have $D(y^{(0)}, \ldots, y^{(T-1)}) = (x^{(0)}, Bw^{(0)}, \ldots, Bw^{(T-2)})$.

The previous discussion leads to the following proposition which gives a characterization of the resilience of a linear control system to attacks on sensors and actuators (similar characterizations with the same flavor have already appeared in the related works [12], [5]):

Proposition 3: Let a control system of the form (5) be given. The following are equivalent:
(i) The system is not resilient against q attacks after T steps
(ii) There exists $x \neq 0$, and vectors $w^{(0)}, \ldots, w^{(T-2)}$ and $e^{(0)}, \ldots, e^{(T-1)}$ with $|\mathsf{supp}(w^{(0)}) \cup \cdots \cup \mathsf{supp}(w^{(T-2)})| + |\mathsf{supp}(e^{(0)}) \cup \cdots \cup \mathsf{supp}(e^{(T-1)})| \leq 2q$ such that

$$\mathcal{O}_T x^{(0)} + \mathcal{M}_T \begin{bmatrix} Bw^{(0)} \\ \vdots \\ Bw^{(T-2)} \end{bmatrix} + \begin{bmatrix} e^{(0)} \\ \vdots \\ e^{(T-1)} \end{bmatrix} = 0 \in \mathbb{R}^{pT}$$

A. Decoding using optimization

In this section we consider the problem of designing a decoding algorithm that recovers the sequence of states despite attacks on sensors and actuators. Like in our previous work [1], one can formulate the decoding problem when there are attacks on both sensors and actuators as an optimization problem. Indeed assume we have received measurements

[2] Actually the assumption —that q errors are not correctable after n steps— only justifies the existence of such attack sequences up to $t = n$ (and not for all t), but using the Cayley-Hamilton theorem one can extend the sequences $e_a^{(t)}$ and $e_b^{(t)}$ to infinite sequences for all t.

[3] In fact the condition asks for the reconstruction of $(x^{(0)}, Bw^{(0)}, \ldots, Bw^{(T-2)})$ only and not the attack vectors $(e^{(t)})$. However it is easy to see that, for the system (5), if we can reconstruct $x^{(0)}, Bw^{(0)}, \ldots, Bw^{(T-2)}$, then we can also reconstruct the attack vectors $e^{(0)}, \ldots, e^{(T-1)}$.

$y^{(0)}, \dots, y^{(T-1)}$ and that we wish to reconstruct the sequence of states $x^{(0)}, \dots, x^{(T-1)}$. Then this can be done by solving the following optimization problem:

$$\begin{array}{ll} \text{minimize} & |\hat{K}| + |\hat{L}| \\ \text{subject to} & \mathsf{supp}(\hat{e}^{(t)}) \subseteq \hat{K}, \mathsf{supp}(\hat{w}^{(t)}) \subseteq \hat{L} \\ & y^{(t)} = C\hat{x}^{(t)} + \hat{e}^{(t)} \\ & \hat{x}^{(t+1)} = A\hat{x}^{(t)} + B(u^{(t)} + \hat{w}^{(t)}) \end{array} \qquad (6)$$

The optimization variables are indicated by a "hat" (e.g., $\hat{x}^{(t)}$, etc.); the other variables (namely, $y^{(t)}$ and $u^{(t)}$) are given. The optimization program above finds the *simplest* possible explanation to the received data $y^{(0)}, \dots, y^{(T-1)}$, i.e., the one with the smallest number of attacked nodes. One can easily show that if the system is resilient against q attacks (in the sense of definition 3), and if the number of actual attacks is less than q, then the output of the optimization problem above gives the correct sequence of states, i.e., $\hat{x}^{(0)} = x^{(0)}, \dots, \hat{x}^{(T-1)} = x^{(T-1)}$.

Unfortunately though, it is known that solving this optimization problem is hard in general [1]. In our previous work we used ideas from compressed sensing and error correction over the reals [10] to relax the decoder by replacing the "ℓ_0" norm (that measures the *cardinality* of the attack set) by an ℓ_1 norm. This relaxation can also be done here when considering attacks on actuators in addition to attacks on sensors and it leads to the following tractable decoder which can be solved efficiently using standard convex optimization software such as [13]:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{p} \|\hat{E}_i\|_{\ell_r} + \lambda \sum_{i=1}^{m} \|\hat{W}_i\|_{\ell_r} \\ \text{subject to} & \hat{E}_i = (\hat{e}_i^{(0)}, \dots, \hat{e}_i^{(T-1)}) \\ & \hat{W}_i = (\hat{w}_i^{(0)}, \dots, \hat{w}_i^{(T-2)}) \\ & y^{(t)} = C\hat{x}^{(t)} + \hat{e}^{(t)} \\ & \hat{x}^{(t+1)} = A\hat{x}^{(t)} + B(u^{(t)} + \hat{w}^{(t)}) \end{array} \qquad (7)$$

For each i the auxiliary variables $\hat{E}_i \in \mathbb{R}^T$ and $\hat{W}_i \in \mathbb{R}^T$ carry the i'th components of the attack vectors over the time horizon $t = 0, \dots, T-1$ (cf. constraints of the optimization program). Thus if $\|\hat{E}_i\|_{\ell_r} = 0$ then $\hat{e}_i^{(t)} = 0$ for all $t = 0, \dots, T-1$ and the i'th sensor is not attacked, and similarly if $\|\hat{W}_i\|_{\ell_r} = 0$ then the i'th actuator is not attacked. Now observe that the objective function $\sum_{i=1}^{p} \|\hat{E}_i\|_{\ell_r} + \lambda \sum_{i=1}^{m} \|\hat{W}_i\|_{\ell_r}$ is nothing but a weighted sum of the ℓ_1 norms of the vectors $(\|\hat{E}_i\|_{\ell_r})_{i=1,\dots,p} \in \mathbb{R}^p$ and $(\|\hat{W}_i\|_{\ell_r})_{i=1,\dots,m} \in \mathbb{R}^m$. Note that we have introduced a tuning parameter λ to control the relative weight between the term corresponding to the attacks on sensors and the term corresponding to the attacks on actuators.

To illustrate the behavior of the ℓ_1 decoder, we tested it on a synthetic randomly-generated system with $n = 15$ states, $m = 10$ actuators and $p = 10$ sensors. We used the parameters $\ell_r = \ell_2$ and $\lambda = 10$, and the optimization problem was solved using the software CVX [13]. For different values of $|K|$ (number of attacked sensors) and $|L|$ (number of attacked actuators), we ran the decoder on 200 different initial conditions and attack sets and we recorded the success rate of the decoder. In figure 2 we see that the decoder succeeded in recovering the state of the system despite the attacks when the number of attacked sensors and actuators is small enough.

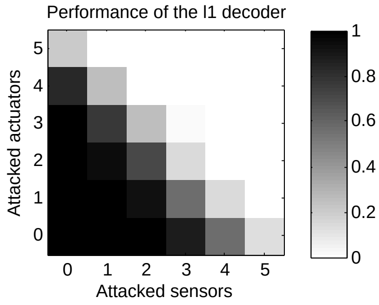

Fig. 2. Performance of the ℓ_1/ℓ_r decoder (7) (with $\ell_r = \ell_2$ and constant $\lambda = 10$) on a random system with $n = 15$ states, $m = 10$ actuators and $p = 10$ sensors. Dark color indicates a high success rate and white color indicates a low success rate. We observe that when the number of attacked sensors and actuators is small enough, the decoder (7) succeeds in recovering the state despite the attacks.

As noted above the ℓ_1 decoder of equation (7) depends on a tuning parameter λ. For the example above we empirically found the value $\lambda = 10$ to be suitable for the considered system. It would be interesting however to see if there is a simple way to directly find the best value of λ from the data and the parameters of the system.

References

[1] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure state-estimation for dynamical systems under active adversaries," in *49th Annual Allerton Conference on Communication, Control, and Computing, 2011*, 2011.

[2] A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems," in *Proceedings of the 3rd conference on Hot topics in security*. USENIX Association, 2008, p. 6.

[3] A. Teixeira, S. Amin, H. Sandberg, K. Johansson, and S. Sastry, "Cyber security analysis of state estimators in electric power systems," in *49th IEEE Conference on Decision and Control (CDC 2010)*.

[4] O. Kosut, L. Jia, R. Thomas, and L. Tong, "On malicious data attacks on power system state estimation," in *Universities Power Engineering Conference (UPEC), 2010 45th International*. IEEE, 2010, pp. 1–6.

[5] F. Pasqualetti, F. Dörfler, and F. Bullo, "Cyber-physical attacks in power networks: Models, fundamental limitations and monitor design," *arXiv:1103.2795*, 2011.

[6] A. Gupta, C. Langbort, and T. Basar, "Optimal control in the presence of an intelligent jammer with limited actions," in *49th IEEE Conference on Decision and Control (CDC 2010)*, pp. 1096–1101.

[7] S. Sundaram, M. Pajic, C. Hadjicostis, R. Mangharam, and G. Pappas, "The wireless control network: monitoring for malicious behavior," in *49th IEEE Conference on Decision and Control (CDC 2010)*.

[8] Y. Mo, T.-J. Kim, K. Brancik, D. Dickinson, H. Lee, A. Perrig, and B. Sinopoli, "Cyber-physical security of a smart grid infrastructure," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 195–209, 2012.

[9] M. Pajic, S. Sundaram, J. Le Ny, G. Pappas, and R. Mangharam, "The wireless control network: Synthesis and robustness," in *49th IEEE Conference on Decision and Control (CDC 2010)*, pp. 7576–7581.

[10] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, 2005.

[11] P. Antsaklis and A. Michel, *Linear systems*. Birkhauser, 2005.

[12] S. Sundaram and C. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *Automatic Control, IEEE Transactions on*, July 2011, to appear.

[13] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," http://cvxr.com/cvx, Apr. 2011.