Resource Constrained Sensor Attacks by Minimizing Fisher Information

Ingvar Ziemann, Henrik Sandberg

Resource Constrained Sensor Attacks by Minimizing Fisher Information

Ingvar Ziemann, Henrik Sandberg

***Abstract*— We analyze the impact of sensor attacks on a linear state estimation problem subject to variance and sparsity constraints. We show that the maximum impact in a leader-follower game where the attacker first chooses the distribution of an adversarial perturbation and the defender follows by choosing an estimator is characterized by a minimum Fisher information principle. In general, this is a nonlinear variational problem, but we show that it can be reduced to a finite-dimensional mixed integer SDP.**

I. Introduction

One of the main challenges of designing secure cyber-physical systems (CPS) is related to analyzing the risk or potential damage an adversary can cause and how to maintain good operating performance in the presence of adversarial attacks. This is often referred to as *impact analysis* [1]. One such class of attacks correspond to the adversary manipulating sensor outputs, to potentially cause harm down the line by introducing misinformation. To this end, we will introduce a new type of resource limited adversarial attack based on minimizing Fisher information and analyze its impact on state reconstruction performance. We show that optimizing this information quantity is equivalent to a leader-follower game, in which the adversary first selects the distribution of their perturbation after which the defender chooses an estimator. However, the minimization of Fisher information[1] is in general a non-convex infinite-dimensional problem. To resolve this, we present a finite-dimensional reduction which involves solving either a sequence of Semidefinite Programs (SDP) or a single mixed-integer SDP.

a) Related Work: Closely related to our work is the impact assessment problem considered in [2] and [3] in which the authors treat a deterministic maximum impact attack to linear systems subject to 2-norm constraints. Interestingly, the authors of [2] also incorporate sparsity constraints in their analysis which relate to compressive sensing type of arguments [4], for instance by introduction of the so-called *security index*, [5]. Another line of work focuses on stealthy maximum impact attacks by incorporating detector threshholds, see for instance [6], [7], [8] and [9]. Particularly pertintent here are the results of [8] which similarly take an operational impact perspective and consider down-the-line control performance whereas we consider estimation performance. Due to our emphasis on estimation performance and Fisher information, we are also able to draw on well-established statistical principles such as from robust estimation [10], [11] and [12]. Similar models can also be found in the literature on minimax estimation [13], [14] or in the compressive sensing literature, see e.g. [15] for a (Bayesian) model similar to ours.

b) Contribution: With this in mind, our work aims to expand on the existing secure CPS literature, and also bring closer together this branch of work with classical statistical principles. Specifically:

- We characterize in terms of a mixed integer SDP a worst case adversarial attack subject to variance and cardinality constraints and show that this attack minimizes a certain information functional.
- Further, we derive a matching estimator which together with the worst case attack constitute a solution of a leader-follower game in which the adversary first selects a distribution of their attack after which the defender selects an estimator.
- On a more technical note, our SDP characterization of the distribution optimizing the trace of inverse Fisher information may also be of independent interest, as this quantity also plays a key role in certain privacy-preserving mechanisms, e.g. in [16],[17],[18]. By contrast, previous results were only able to characterize this distribution for scalar random variables, [10], or considered relaxed problems [16].

An interesting feature of our results is also that while there exists a solution to the leader-follower game, the derived estimator will in general not be robust (minimax) in the presence of sparsity constraints. In this way, our results can be understood as a lower bound on the impact an adversary can achieve.

c) Notation: For any integer i, we write $[i] = \{1, \dots i\}$. If $S \subset [i]$, $S^\perp$ denotes the complement of S in $[i]$. We use $\succeq$ (and $\succ$) for (strict) inequality in the matrix positive definite partial order. By $\|\cdot\|$ we denote the standard 2-norm and by $\|\cdot\|_\infty$ the matrix operator norm. For any positive semi-definite matrix Q, we write $\|\cdot\|_Q$ for the 2-(semi-)norm weighted by Q and $\dagger$ is used to denote the Moore-Penrose pseudoinverse. We denote $\{e_i\}$ the standard basis in Euclidean space. Further, we write $\mathbf{E}$ for the expectation operator and $*$ for the convolution between two (probability) measures. For derivatives, we use ∂_x for the transpose of the gradient with respect to the vector x; i.e. ∂_x is used to represent a column vector of first derivative operators. All (in)equalities not under an expectation are in this paper taken to hold almost surely.

Ingvar Ziemann (ziemann@kth.se) and Henrik Sandberg (hsan@kth.se) are with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden.

This work was supported in part by the Swedish Research Council (grant 2016-00861), and the Swedish Foundation for Strategic Research (Project CLAS).

[1]To be precise, the maximization of the trace of the inverse of Fisher information.

II. Problem Formulation

We suppose $x \in \mathbb{R}^n$ is a deterministic unknown variable to be estimated. The operator only has access to imperfect observations, $y \in \mathbb{R}^p$ given by the measurement equation

$$y = Hx + Fa + w \tag{1}$$

where $a \in \mathbb{R}^m$ is an adversarial perturbation which has distribution μ_a, $H \in \mathbb{R}^{p\times n}, F \in \mathbb{R}^{p\times m}$ and where $w \in \mathbb{R}^p$ is an $N(0, \Sigma_w)$ disturbance with $\Sigma_w \succ 0$.

The adversary is assumed resource constrained in the sense that

- A1. The adversary can only impact s sensors at once, $(Fa)_i \neq 0$ for at most s indices. To be precise, the distribution of $(Fa)_i$ is supported by $\mathrm{span}\{e_i\}, \mathrm{card}\{e_i\} \leq s$.
- A2. The adversary has an energy constraint $\mathbf{E}a^\top Ca \leq \sigma^2$, $\sigma^2 \in \mathbb{R}, C \succ 0$.

We suppose the goal of the operator is to construct an estimator $\hat{x} = \hat{x}(y)$ of x, such that the statistical risk

$$\mathbf{E}\|x - \hat{x}\|^2 = \mathrm{tr}\,\mathbf{E}(x - \hat{x})(x - \hat{x})^\top \tag{2}$$

is minimized. On the other hand, the adversary seeks to maximize this quantity. To make the problem nontrivial, we need to rule out estimators of the form $\hat{x} = x$ independent from the observed y:

- A3. The estimator $\hat{x}$ of x is unbiased; that is $\mathbf{E}\hat{x} = x$ and its risk is given by (2).
- A4. It is assumed throughout that the unknown variable x is observable through H; $x \perp \ker H$.
- A4'. H has full column rank.

Remark 2.1: We note that A4 can be replaced by A4' without loss of generality. To see this, let P be the projection onto $(\ker H)^\perp = \mathrm{im}(H^\top)$. Note that $PP^\top = I_{\dim \mathrm{im}\, H^\top}$ and $P^\top P$ acts as the identity on the subspace $(\ker H)^\perp$. Since $x = P^\top Px$, by redefining variables as $H_{red} = HP^\top$ and $x_{red} = Px$ we obtain an equivalent measurement system $y = H_{red}x_{red} + Fa + w$.

By virtue of this remark, it suffices to prove all results that follow under A4', noting that the simple transformation described above extends them also to A4.

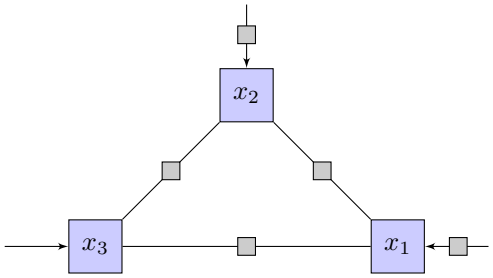
We will consider this problem mainly from the adversary's perspective. That is, we consider a leader-follower game where the adversary goes first and selects the distributions μ_a satisfying A1 and A2 upon which the operator selects an estimator satisfying A3. This game is equivalent to

$$\max_{\mu_a} \min_{\hat{x}} \mathrm{tr}\,\mathbf{E}(x - \hat{x})(x - \hat{x})^\top$$

subject to A1-A3, A4'. Since we always have $\max\min \leq \min\max$, it follows that our game can be interpreted as a *lower bound* on the maximum impact an adversary can inflict in terms of estimation performance. Note also that, since the adversary goes first, no deterministic attack a_d can be optimal, as this essentially corresponds to the strategy δ_{a_d} where δ_{a_d} is a point mass at a_d. Since the defender knows the distribution δ_{a_d}, they can simply subtract a_d from their measurement.

Before proceeding, let us consider an example motivating the problem's relevance.

Example 2.2: Let us consider the DC power flow depicted in the figure below. Analysis of such networks is commonplace in the power systems state estimation literature, [19].



Suppose the system operator wishes to reconstruct the voltage phase angles at x_1, x_2, x_3 and that noisy active power flow measurements are available at the grey squares. This can be formulated as $y, w \in \mathbb{R}^5, x \in \mathbb{R}^3$ with noisy measurements, so that $y = Hx + Fa + w$ with

$$H = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \\ 2 & -1 & -1 \\ -1 & 2 & -1 \end{bmatrix}, \qquad \Sigma_w = I_5.$$

Let us assume further that the attacker has possibility to corrupt one of these measurements, so that $a \in \mathbb{R}^5$, $F = I_5$ and that the adversary's constraints are $\mathbf{E}a^\top a = 1$ and $\mathrm{card}\{a_i \neq 0\} \leq 1$ (i.e., $s = 1$). We return to, and solve, this example in Section IV.

Example 2.3: Consider a linear dynamical system

$$X_{t+1} = H_tX_t + B_tA_t + W_t, \quad Y_t = CX_t, \quad t = 1, \ldots, T$$

with unknown initial condition $X_0 = x$. This can also be handeled by the model (1), simply by converting them into Toeplitz form, i.e. writing $y = (Y_1, \ldots, Y_T)$ as a linear combination of $a = (A_1, \ldots, A_t)$, $w = (W_1, \ldots, W_T)$.

a) Operational Interpretation of the Constraints: From an operational point of view, the constraint A1 is rather it clear; it signifies the belief that the adversary has limited resources in terms of the number of sensors they can corrupt. However, the operator is not sure precisely which these are. Constraint A2, on the other hand, deserves further justification. More specifically, one may ask why an adversary launching a cyber-attack would have a seemingly physical constraint on its input energy. To give A2 an operational justification, suppose that the system operator runs an anomaly detection scheme such as a chi-squared detector, which is commonly considered in the literature, e.g. [7]. In this case, the covariance of a directly impacts the detection probability of the attack and A2 can thus be understood as to embody the adversary's desire to remain undetected (with some probability depending on C). As for A3, it is merely one way to ensure that estimators of the form $\hat{x} = x$ are inadmissible. It is not hard to see that our results remain valid if one removes A3 and instead replaces (2)

with $\sup_x \operatorname{tr} \mathbf{E}(x-\hat{x})(x-\hat{x})^\top$ instead of the unbiasedness assumption. The final assumption A4 is necessary for the existence of an unbiased estimator. Without it, x cannot in general reconstructed without extra side information.

III. Least Informative Distributions

We begin our process toward finding a solution of the game under A1-A3 and A4' by considering certain estimation-theoretic principles. We will consider attacks which are *least informative about* x in a certain sense, and then later show that these attacks actually are part of the required saddle-point.

It is a classical result in statistics that for unbiased estimators, it holds true that

$$\mathbf{E}(x-\hat{x})(x-\hat{x})^\top \succeq J_x^\dagger$$

where $J_x(p)$ is the Fisher information (matrix). This inequlity is known as the Cramér-Rao bound.

Definition 3.1: Suppose that $x \in \Theta$ where Θ is an open subset of $\mathbb{R}^n$. Let $\{p(y;x)\}$ be a family of densities over $\mathbb{R}^m$ depending smoothly on x. Then Fisher information (matrix) is defined as

$$J_x(p) = \int_{\mathbb{R}^p} \partial_x \log p(y;x) [\partial_x \log p(y;x)]^\top p(y;x) dy \quad (3)$$

whenever the integral exists.

This note investigates a class of attacks which are *least informative about* x in the sense that the strategy of the adversary is characterized by the distribution of the random variable which is worst possible in terms of the bound (3). With this in mind, we now seek to characterize the distributions which maximize the trace of $J_x^\dagger(p)$, where, in our model, $p(y;x)$ corresponds to the density of y in (1).

Definition 3.2: If μ_a is the probability measure of a in (1) and $p(y;x)$ corresponds to the density of y in (1), we say that μ_a is *least informative about* x if it maximizes $\operatorname{tr}(J_x(p))^\dagger$ subject to the constraints A1 and A2.

Observe that a priori, we cannot guarantee the existence of the integral (3). The next lemma shows that this is well-defined.

Lemma 3.3: For any non-degenerate normal random variable w, the probability measure of the random variable $Fa+w$ admits a C^∞ density, say q, with respect to Lebesque measure.

Proof: Let μ_{Fa} be the probability measure of the variable Fa and γ_{Σ_w} be the normal measure with covariance Σ_w. The distribution of the $Fa+w$ is given by the $\mu_{Fa} * \gamma_{\Sigma_w}$ which has a C^∞ density with respect to Lebesque measure by a well-known property of Gaussian convolutions since γ_{Σ_w} has a C^∞ density. ∎

Since $Fa+w$ admits a C^∞ density q, Fisher information of the model (1) exists and can be further simplified.

Lemma 3.4: Let q be the density of $Fa+w$. Fisher information of the model (1) is given by

$$J_x(p) = H^\top J(q) H \quad (4)$$

where

$$J(q) = \int_{\mathbb{R}^p} \partial_x \log q(v) [\partial_x \log q(v)]^\top q(v) dv. \quad (5)$$

Proof: Note that $p(y;x) = q(y-Hx)$, from which it follows that

$$\begin{aligned} J_x(p) &= \int_{\mathbb{R}^p} \partial_x \log p(y;x) [\partial_x \log p(y;x)]^\top p(y;x) dy \\ &= \int_{\mathbb{R}^p} \partial_x \log q(y-Hx) [\partial_x \log q(y-Hx))]^\top q(y-Hx) dy \\ &= \int_{\mathbb{R}^p} H^\top \partial_x \log q(v) [\partial_x \log q(v)]^\top H q(v) dv. \end{aligned}$$

∎

The least informative distribution is thus given by the following optimization problem.

$$\begin{cases} \sup_{\mu_a} & \operatorname{tr}(H^\top J(q) H)^\dagger \\ \text{s.t.} & \mathbf{E} a^\top C a \le \sigma^2 \\ & q = \mu_{Fa} * \gamma_{\Sigma_w} \\ & \text{the support of the distribution of } Fa \\ & \text{satisfies } (Fa)_i \neq 0 \text{ for at most } s \text{ indices} \end{cases} \quad (6)$$

where μ_a is the probability measure of a.

There are a three major issues with Problem (6). First, the mapping $q \mapsto \operatorname{tr}(H^\top J(q) H)^\dagger$ is non-convex, and second, even if there were no further constraints, this objective is non-linear and infinite-dimensional making it rather hard to solve for. Moreover, there is a third issue relating to the sparsity constraint A1 which also breaks convexity of the problem. We will now turn to reducing and relaxing to a finite-dimensional mixed integer semidefinite program.

a) Reduction to a Finite-Dimensional Convex Objective: Our reduction relies on arguments due to [12] which show that Fisher information of any probability model is lower-bounded (in semidefinite order) by a Gaussian model with the same mean and covariance matrix. Since this Gaussian model is admissible as a strategy for the adversary, we can restrict the optimization in Problem (6) to Gaussian measures, which (if mean zero) are completely characterized by their covariance matrices. Observe also that if q is Gaussian with covariance Σ_q, we have $J_q = \Sigma_q^\dagger$, where J_q is given by (5). These observations yield the following lemma.

Lemma 3.5: Problem (6) is equivalent to

$$\begin{cases} \max_{\Sigma_a \succeq 0} & \operatorname{tr}\left(H_{\operatorname{im} H}^{-2} (\tilde{\Sigma}_{11} - \tilde{\Sigma}_{12} \tilde{\Sigma}_{22}^{-1} \tilde{\Sigma}_{12}^\top) \right) \\ \text{s.t.} & \operatorname{tr} \Sigma_a C \le \sigma^2 \\ & \tilde{\Sigma}_{Fa+w} = \tilde{\Sigma}_{Fa} + \tilde{\Sigma}_w \\ & \text{the support of the distribution of } Fa \\ & \text{satisfies } (Fa)_i \neq 0 \text{ for at most } s \text{ indices} \end{cases} \quad (7)$$

where $U, H_{\operatorname{im} H}$ and $\tilde{\Sigma}_{Fa+w}$ are given by (8) and (9). In this case, the least informative distribution is $N(0, \Sigma_a^*)$ where Σ_a^* solves the problem.

Proof: By the results of [12] we have

$$H^\top J(q) H \succeq H^\top \Sigma_{Fa+w}^{-1} H$$

and equality holds if the distribution of $Fa+w$ is normal. We may thus restrict the domain of the objective to mean zero Gaussian variables. Moreover, by the chain rule for

Fisher information, we may further restrict attention to those a which are independent of w. In this case, the objective becomes

$$\operatorname{tr}(H^\top \Sigma_{Fa+w}^{-1} H)^{-1}$$

where the inverses exist due the positive-definiteness of Σ_w and the full column rank of H. Let us now change coordinates by introducing the range-nullspace SVD of H as

$$H = \begin{bmatrix} U_{\operatorname{im} H} & U_{\ker H^\top} \end{bmatrix} \begin{bmatrix} H_{\operatorname{im} H} \\ 0 \end{bmatrix} V_{\operatorname{im} H^\top}, \tag{8}$$

or simply $H = U H_s V$. Now let

$$\tilde{\Sigma}_{Fa+w} = \begin{bmatrix} \tilde{\Sigma}_{11} & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{21} & \tilde{\Sigma}_{22} \end{bmatrix} = U \Sigma_{Fa+w} U^\top \tag{9}$$

with $\tilde{\Sigma}_{11} \in \mathbb{R}^{n \times n}$ and $\tilde{\Sigma}_{22} \in \mathbb{R}^{(p-n)\times(p-n)}$. Then, if we denote the block-entries of $\tilde{\Sigma}^{-1}$ as $\tilde{\Sigma}_{ij}^{inv}$,

$$\begin{aligned}
(H^\top \Sigma_{Fa+w}^{-1} H)^{-1} &= \Bigg(V_{\operatorname{im} H^\top}^\top \begin{bmatrix} H_{\operatorname{im} H} & 0 \end{bmatrix} \begin{bmatrix} U_{\operatorname{im} H}^\top \\ U_{\ker H^\top}^\top \end{bmatrix} \\
&\quad \times \Sigma_{Fa+w}^{-1} \begin{bmatrix} U_{\operatorname{im} H} & U_{\ker H^\top} \end{bmatrix} \begin{bmatrix} H_{\operatorname{im} H} \\ 0 \end{bmatrix} V_{\operatorname{im} H^\top} \Bigg)^{-1} \\
&= \Bigg(V_{\operatorname{im} H^\top}^\top \begin{bmatrix} H_{\operatorname{im} H} & 0 \end{bmatrix} \begin{bmatrix} \tilde{\Sigma}_{11}^{inv} & \tilde{\Sigma}_{12}^{inv} \\ \tilde{\Sigma}_{21}^{inv} & \tilde{\Sigma}_{22}^{inv} \end{bmatrix} \begin{bmatrix} H_{\operatorname{im} H} \\ 0 \end{bmatrix} V_{\operatorname{im} H^\top} \Bigg)^{-1} \\
&= \left(V_{\operatorname{im} H^\top}^\top H_{\operatorname{im} H} \tilde{\Sigma}_{11}^{inv} H_{\operatorname{im} H} V_{\operatorname{im} H^\top} \right)^{-1}.
\end{aligned}$$

Since further $\tilde{\Sigma}_{11}^{inv} = (\tilde{\Sigma}_{11} - \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}\tilde{\Sigma}_{12}^\top)^{-1}$ by the Schur complement formula (since $\tilde{\Sigma} \succ 0$, these inverses exist), we have that

$$\begin{aligned}
&\operatorname{tr}(H^\top \Sigma_{Fa+w}^{-1} H)^{-1} \\
&\qquad = \operatorname{tr}\left(H_{\operatorname{im} H}^{-1} (\tilde{\Sigma}_{11} - \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}\tilde{\Sigma}_{12}^\top) H_{\operatorname{im} H}^{-1} \right).
\end{aligned} \tag{10}$$

Observing that $\Sigma_{Fa+w} = F\Sigma_a F^\top + \Sigma_w$ by independence and if we let $\tilde{\Sigma}_{Fa} = U\Sigma_{Fa}U^\top, \tilde{\Sigma}_w = U\Sigma_w U^\top$ we also have $\tilde{\Sigma}_{Fa+w} = \tilde{\Sigma}_{Fa} + \tilde{\Sigma}_w$. ∎

Remark 3.6: The change of coordinates (8) induces the equivalent measurement equation

$$\tilde{y} = U^{-1} y = \begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \end{bmatrix} = \begin{bmatrix} H_{\operatorname{im} H} \\ 0 \end{bmatrix} \tilde{x} + U^{-1} F a + U^{-1} w$$

where $\tilde{x} = V_{\operatorname{im} H^\top} x$.

b) Reduction to Mixed Integer SDP: Problem (7) still has awkward dependencies through $\tilde{\Sigma}_{22}^{-1}$ and the coupling $\tilde{\Sigma}_{Fa+w} = \tilde{\Sigma}_{Fa} + \tilde{\Sigma}_w$. Our first theorem shows that this can helped and that the least informative distribution for our problem can be found by solving a number of semidefinite programs.

Theorem 3.7: Problem (6) is equivalent to the mixed integer semidefinite program

$$\begin{cases}
\max\limits_{\substack{R \succeq 0, \\ \tilde{\Sigma}_{Fa} \succeq 0, \\ S^\perp \subset [p]}} & \operatorname{tr}\left(H_{\operatorname{im} H}^{-2} R \right) \\
\text{s. t.} & \begin{bmatrix} \tilde{\Sigma}_{Fa,11} + \tilde{\Sigma}_{w,11} - R & \tilde{\Sigma}_{Fa,12} + \tilde{\Sigma}_{w,12} \\ \tilde{\Sigma}_{Fa,12}^\top + \tilde{\Sigma}_{w,12}^\top & \tilde{\Sigma}_{Fa,22} + \tilde{\Sigma}_{w,22} \end{bmatrix} \succeq 0 \\
& \operatorname{tr} \tilde{\Sigma}_{Fa} U F^{\dagger,\top} C F^\dagger U^\top \leq \sigma^2 \\
& \operatorname{tr} \tilde{\Sigma}_{Fa} U (I - FF^\dagger) U^\top = 0 \\
& \operatorname{tr} e_i^\top U^\top \tilde{\Sigma}_{Fa} U e_i = 0, \forall i \in S^\perp \\
& |S^\perp| = p - s
\end{cases} \tag{11}$$

where $U, H_{\operatorname{im} H}$ and $\tilde{\Sigma}_w$ are given by (8) and (9). In this case, the least informative distribution is $N(0, \tilde{F}^\dagger U^\top \tilde{\Sigma}_{Fa}^* U F^{\dagger\top})$ where $\tilde{\Sigma}_{Fa}^*$ solves the problem.

Proof: We replace the Schur complement in the objective with a variable, R. Let $R \preceq (\tilde{\Sigma}_{11} - \tilde{\Sigma}_{12}\tilde{\Sigma}_{22}^{-1}\tilde{\Sigma}_{12}^\top)$, which can be written as the LMI

$$\begin{bmatrix} \tilde{\Sigma}_{11} - R & \tilde{\Sigma}_{12} \\ \tilde{\Sigma}_{12}^\top & \tilde{\Sigma}_{22} \end{bmatrix} \succeq 0$$

which, since $\tilde{\Sigma}_{Fa+w} = \tilde{\Sigma}_{Fa} + \tilde{\Sigma}_w$ is equivalent to

$$\begin{bmatrix} \tilde{\Sigma}_{Fa,11} + \tilde{\Sigma}_{w,11} - R & \tilde{\Sigma}_{Fa,12} + \tilde{\Sigma}_{w,12} \\ \tilde{\Sigma}_{Fa,12}^\top \tilde{\Sigma}_{w,12}^\top & \tilde{\Sigma}_{Fa,22} + \tilde{\Sigma}_{w,22} \end{bmatrix} \succeq 0$$

and thus yields the desired LMI.

Next, we need to convert $\operatorname{tr} \Sigma_a C \leq \sigma^2$ into the variable $\tilde{\Sigma}_{Fa}$. However, since F has full column rank,

$$\Sigma_a = F^\dagger F \Sigma_a F^\top F^{\dagger\top} = F^\dagger \Sigma_{Fa} F^{\dagger\top} = F^\dagger U^\top \tilde{\Sigma}_{Fa} U F^{\dagger\top}.$$

Moreover, the constraint $\tilde{\Sigma}_{Fa+w} = \tilde{\Sigma}_{Fa} + \tilde{\Sigma}_w$ can be eliminated by requiring that Σ_{Fa} is in the range of the mapping $\Sigma \mapsto F\Sigma F^\top$, which holds if and only if $\ker \Sigma_{Fa} \subseteq \ker F^\top$. In turn, this condition is equivalent to $\operatorname{tr} \Sigma_{Fa}(I - FF^\dagger) = 0$, or $\operatorname{tr} U^\top \tilde{\Sigma}_{Fa} U (I - FF^\dagger) = 0$, since $I - FF^\dagger$ is the orthogonal projector onto the kernel of $F^\top$ and since the variable $\tilde{\Sigma}_{Fa}$ is positive definite.

Finally, the constraint $(Fa)_i \neq 0$ for at most s indices can also be written in terms of a sequence of linear inequalities in $\tilde{\Sigma}_{Fa}$. To do so, note first that $(Fa)_i \neq 0$ for at most s indices is equivalent to $(Fa)_i = 0$ for at least $p - s$ indices. Moreover, $(Fa)_i = e_i^\top Fa$ where we recall that e_i denotes i:th standard basis vector for $\mathbb{R}^p$. Now, $e_i^\top Fa = 0$ almost surely if and only if

$$\begin{aligned}
0 = \mathbf{E} \operatorname{tr}(e_i^\top Fa)(e_i^\top Fa)^\top &= \operatorname{tr} e_i^\top F \mathbf{E} a a^\top F e_i \\
&= \operatorname{tr} e_i^\top F \mathbf{E} \Sigma_a F^\top e_i = \operatorname{tr} e_i^\top \Sigma_{Fa} e_i \\
&= \operatorname{tr} e_i^\top U^\top \tilde{\Sigma}_{Fa} U e_i.
\end{aligned}$$

Hence $(Fa)_i \neq 0$ for at most s indices if and only if $\operatorname{tr} e_i^\top U^\top \tilde{\Sigma}_{Fa} U e_i = 0$ for all $i \in S^\perp$ for some $S^\perp \subset [p]$ and $|S^\perp| = p - s$. ∎

IV. Game-Theoretic Interpretation

Returning to the leader-follower game introduced in Section II, we now show that it is optimal for the adversary to select $\mu_a = N(0, \Sigma_a^*)$ where $\Sigma_a^* = \tilde{F}^\dagger U^\top \tilde{\Sigma}_{Fa}^* U F^{\dagger,\top}$ is a solution of Theorem 3.7.

Theorem 4.1: For any choice of μ_a satisfying A1 and A2 and any unbiased estimator (A3) $\hat{x}$ of x satisfying A4 we have that

$$\max_{\mu_a} \min_{\hat{x}} \operatorname{tr} \mathbf{E}(x - \hat{x})(x - \hat{x})^\top \tag{12}$$

is given by (7). In addition, any solution of Theorem 3.7 yields the optimal μ_a of (12) and the optimal $\hat{x}$ is given by

$$\hat{x} = (H^\top \Sigma_{Fa+w} H)^{-1} H^\top \Sigma_{Fa+w}^{-1} y \tag{13}$$

where Σ_{Fa+w} is the covariance of $Fa + w$.

Proof: Let $\hat{x}$ be given by (13) and observe the chain of inequalities

$$\mathbf{E}(x - \hat{x})(x - \hat{x})^\top = (H^\top \Sigma_{Fa+w}^{-1} H)^{-1} \succeq [J_x(p)]^{-1},$$

where the first equality is a standard result for this choice of $\hat{x}$ (see e.g. [20], Theorem 13.2) and the second inequality follows by Theorem 6.1 (or Cramér-Rao). Hence the error $\operatorname{tr}(H^\top \Sigma_{Fa+w}^{-1} H)^{-1}$ is always attainable, whereas, if a is Gaussian, the second inequality is tight and no better error is attainable. By Theorem 3.7 it follows that (11) gives the value of (12). ∎

Notice that the solution μ_a in general is not unique. To see this, let us reconsider Example 2.2.

Example 4.2 (Example 2.2 continued): Observe that Theorems 3.7 and 4.1 together imply that the solution to the leader follower game with measurement equation

$$y = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \\ 2 & -1 & -1 \\ -1 & 2 & -1 \end{bmatrix} x + a + w \tag{14}$$

with $\mathbf{E}a^\top a \leq 1, \operatorname{card}\{a_i \neq 0\} \leq 1$, and $y, a, w \in \mathbb{R}^5, x \in \mathbb{R}^3$ can be restricted to Gaussian attacks and that the optimal attacks are necessarily of the form $a \sim N(0, E_i)$ where $E_i = (E_{jk}) = (\delta_{jki})$; i.e., the entire perturbation budget is spent on one channel $i \in [5]$. Moreover, since H in (14) does not have full rank, we factor out the nullspace of H as described in Remark 2.1 to form the reduced matrix

$$H_{red} \approx \begin{bmatrix} -1.41 & 0 \\ -0.71 & 1.22 \\ 0.71 & 1.22 \\ -2.12 & 1.22 \\ 2.12 & 1.22 \end{bmatrix}.$$

Concretely, H_{red} is formed using the matlab commands

```
[U S V ]  = svd(H);
P = V(:,1:2);
Hred = H * V(:,1:2);
```

since the third column of `V` formed in the code above spans the nullspace of H.

Hence, we only need to evaluate

$$V_i = \operatorname{tr}(H_{red}^\top (\Sigma + E_i)^{-1} H_{red})^{-1}, \qquad i \in [5]$$

to decide the solution of the game. A straightforward computation shows that $V_1 \approx 0.26, V_2 = V_3 \approx 0.28, V_4 = V_5 \approx 0.3$. Hence, from the adversary's perspective, it is optimal to attack the fourth or fifth measurement.

It is tempting to ascribe robustness properties to the estimator $\hat{x}$ of Theorem 4.1. However, the solution derived does not in general constitute a Nash equilibrium. We illustrate this by the following example.

Example 4.3: Suppose $y, x, a, w \in \mathbb{R}^2$ with

$$y = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} x + a + w, \qquad \Sigma_w = I_2$$

and that the adversary has the constraints $\mathbf{E}a^\top a \leq 1, \operatorname{card}\{a_i \neq 0\} \leq 1$. As in Example 4.2 we only need to compute

$$V_i = \operatorname{tr}(H^\top (\Sigma_w + E_i)^{-1} H)^{-1} = 5/3, \qquad i \in [2].$$

Suppose that the attacker plays $N(0, E_1)$ and that the defending player chooses the estimator $\hat{x} = (H^\top (\Sigma_w + E_1) H)^{-1} H^\top (\Sigma_w + E_1)^{-1} y$ in this case, as computed above, the leader-follower game has value $V_1 = 5/3$. On the other hand, if the adversary deviates and instead plays $N(0, E_2)$, the risk of the estimator $\hat{x} = (H^\top (\Sigma_w + E_1) H)^{-1} H^\top (\Sigma_w + E_1)^{-1} y$ is $\mathbf{E}\|\hat{x} - x\|^2 > 5/3$ (see [20]). Indeed, if the defender goes first, by symmetry, the best they can do is to use the estimator $\hat{x} = (H^\top H)^{-1} H^\top y$ (see [14]).

Example 4.3 shows why Theorem 4.1 cannot in general give us a Nash equilibrium, the reason being that the cardinality constraint on the adversarial perturbation making the strategy set of the adversary non-convex. For instance, in the example above, $N(0, E_1)$ and $N(0, E_2)$ are both admissible, but no convex combination of these measures $\alpha N(0, E_1) + (1 - \alpha) N(0, E_2), \alpha \in (0, 1)$ is admissible. In particular, the usual convexity assumptions needed for minimax theorems do not hold, see e.g. [21]. More practically, in the example above, a minimax defender would need to defend against $\alpha N(0, E_1)$ and $\alpha N(0, E_2)$ attacks simultaneously, and the only reasonable estimator for this is $\hat{x} = (H^\top H)^{-1} H^\top y$, which is sub-optimal for either attack considered in isolation.

V. Conclusion

We introduced a model for sensor attacks on cyber-physical systems based on an estimation objective. It was shown that the maximal impact in the associated leader-follower game is characterized by the distribution minimizing Fisher information and that this distribution can be found by solving a mixed integer SDP.

Future Work: We saw Example 4.3 that the attack-estimator pair which achieves $\max\min$ is not necessarily $\min\max$. Essentially, this is due to the sparsity assumption-which breaks convexity of the problem and thus induces a gap between $\max\min$ and $\min\max$. Our work can in some sense be interpreted as giving a lower bound for the maximum impact, whereas characterizing the $\min\max$, a game

in which the estimator is selected first, could complement this viewpoint to give the corresponding upper bound. In particular, it would be interesting to see if such a solution also has a characterization in terms of Fisher information.

REFERENCES

[1] M. S. Chong, H. Sandberg, and A. M. Teixeira, "A tutorial introduction to security and privacy for cyber-physical systems," in *2019 18th European Control Conference (ECC)*, IEEE, 2019, pp. 968–978.

[2] A. Teixeira, K. C. Sou, H. Sandberg, and K. H. Johansson, "Quantifying cyber-security for networked control systems," in *Control of cyber-physical systems*, Springer, 2013, pp. 123–142.

[3] A. Teixeira, H. Sandberg, and K. H. Johansson, "Strategic stealthy attacks: The output-to-output l2-gain," in *2015 54th IEEE Conference on Decision and Control (CDC)*, IEEE, 2015, pp. 2582–2587.

[4] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[5] J. M. Hendrickx, K. H. Johansson, R. M. Jungers, H. Sandberg, and K. C. Sou, "Efficient computations of a security index for false data attacks in power networks," *IEEE Transactions on Automatic Control*, vol. 59, no. 12, pp. 3194–3208, 2014.

[6] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, pp. 135–148, 2015.

[7] Y. Mo and B. Sinopoli, "On the performance degradation of cyber-physical systems under stealthy integrity attacks," *IEEE Transactions on Automatic Control*, vol. 61, no. 9, pp. 2618–2624, 2015.

[8] Y. Chen, S. Kar, and J. M. Moura, "Optimal attack strategies subject to detection constraints against cyber-physical systems," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1157–1168, 2017.

[9] J. Milošević, H. Sandberg, and K. H. Johansson, "Estimating the impact of cyber-attack strategies for stochastic networked control systems," *IEEE Transactions on Control of Network Systems*, 2019.

[10] P. J. Huber, *Robust Statistics*. John Wiley & Sons, 2004, vol. 523.

[11] P. Stoica and P. Babu, "The gaussian data assumption leads to the largest cramér-rao bound [lecture notes]," *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 132–133, 2011.

[12] M. Stein, A. Mezghani, and J. A. Nossek, "A lower bound for the fisher information measure," *IEEE Signal Processing Letters*, vol. 21, no. 7, pp. 796–799, 2014.

[13] J. Hodges, E. Lehmann, *et al.*, "Some problems in minimax point estimation," *The Annals of Mathematical Statistics*, vol. 21, no. 2, pp. 182–197, 1950.

[14] R. Radner *et al.*, "Minimax estimation for linear regressions," *The Annals of Mathematical Statistics*, vol. 29, no. 4, pp. 1244–1250, 1958.

[15] M. Sundin, S. Chatterjee, and M. Jansson, "Combined modeling of sparse and dense noise improves bayesian rvm," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, IEEE, 2014, pp. 1841–1845.

[16] F. Farokhi and H. Sandberg, "Ensuring privacy with constrained additive noise by minimizing fisher information," *Automatica*, vol. 99, pp. 275–288, 2019.

[17] I. Ziemann and H. Sandberg, "Parameter privacy versus control performance: Fisher information regularized control," 2019.

[18] F. Farokhi, "Privacy-preserving constrained quadratic optimization with fisher information," *IEEE Signal Processing Letters*, vol. 27, pp. 545–549, 2020.

[19] A. Abur and A. G. Exposito, *Power system state estimation: theory and implementation*. CRC press, 2004.

[20] J. R. Magnus and H. Neudecker, *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2019.

[21] M. Sion *et al.*, "On general minimax theorems.," *Pacific Journal of mathematics*, vol. 8, no. 1, pp. 171–176, 1958.

VI. APPENDIX

Covariance bound on Fisher information: We frequently use the following Theorem due to [12].

Theorem 6.1: Let $y \in \mathbb{R}^p$ be a square integrable random variable with parametrized density $p(y;x)$, $x \in \Theta \subset \mathbb{R}^m$. Define

$$\mu(x) = \int yp(y;x)dy,$$

$$\Sigma(x) = \int (y-\mu(x))(y-\mu(x))^\top p(y;x)dy,$$

$$J_x(p) = \int \partial_x \log p(y;x)[\partial_x \log p(y;x)]^\top p(y;x)dy.$$

Then

$$J_x(p) \succeq [\partial_x \mu(x)]^\top \Sigma^{-1}(x)[\partial_x \mu(x)]$$

where the left hand side is interepreted as infinity, if it does not exist.