Robust Detection in the Presence of Integrity Attacks

Yilin Mo*, João Hespanha†, Bruno Sinopoli*

***Abstract*— We consider the estimation of a binary random variable based on m noisy measurements that can be manipulated by an attacker. The attacker is assumed to have full information about the true value of the variable to be estimated and about the value of all the measurements. However, the attacker has limited resources and can only manipulate n of the m measurements. The problem is formulated as a minimax optimization, where one seeks to construct an optimal detector that minimizes the "worst-case" probability of error against all possible manipulations by the attacker. We show that if the attacker can manipulate at least half the measurements ($n \geq m/2$) then the optimal worst-case estimator should ignore *all* m measurements and be based solely on the a-priori information. When the attacker can manipulate less than half the measurements ($n < m/2$), we show that the optimal estimator is a threshold rule based on a Hamming-like distance between the (manipulated) measurement vector and two appropriately defined sets. For the special case where $m = 2n + 1$, our results provide a constructive procedure for the optimal estimator.**

I. Introduction

The increasing use of networked embedded sensors to monitor and control critical infrastructures such as the power grid, transportation systems and built environments provides potential malicious agents with the opportunity to disrupt their operations by corrupting sensor measurements.

Supervisory Control And Data Acquisition (SCADA) systems, for example, implement the distributed control systems that run a wide range of safety critical plants and processes, including manufacturing, water and gas treatment and distribution, facility control and power grids. A successful attack to SCADA systems may significantly hamper the economy, the environment, and may even lead to the loss of human life. The first-ever SCADA system malware (called Stuxnet) was found in July 2010 and rose significant concern about SCADA system security [1], [2]. While SCADA systems are currently mostly isolated, next generation SCADA will make extensive use of widespread sensing and networking, both wired and wireless, making critical infrastructures susceptible to cyber security threats.

This research is supported in part by CyLab at Carnegie Mellon under grant DAAD19-02-1-0389 from the Army Research Office Foundation and grant NGIT2009100109 from Northrop Grumman Information Technology Inc Cybersecurity Consortium. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either express or implied, of ARO, CMU, or the U.S. Government or any of its agencies.

*: Yilin Mo and Bruno Sinopoli are with the ECE department of Carnegie Mellon University, Pittsburgh, PA. Email: ymo@andrew.cmu.edu, brunos@ece.cmu.edu

†: João Hespanha is with the ECE department of University of California, Santa Barbara, CA. Email: hespanha@ece.ucsb.edu

The research community has acknowledged the importance of addressing the challenge of designing secure detection, estimation and control systems [3].

We consider a robust detection problem inspired by security concerns that arise from the possible manipulation of sensor data. We focus our attention on the estimation of a binary random variable θ from independent measurements collected by m sensors, with the caveat that some of these measurements can be manipulated by an attacker. The attacker is assumed to have full information about the true value of θ and all the measurements and uses this information to manipulate the data available to the detector. Limitations in the resources available to the attacker enable him to only manipulate n of the m sensors. However, the attacker has total control over the corrupted sensors, as he can change their values arbitrarily. To minimize the detector's performance degradation in the face of such attacks, we construct minimax detectors that minimize the "worst-case" probability of detection error, where worst-case refers to all possible manipulations available to the attacker.

We start by considering the case $n \geq m/2$, in which the attacker can manipulate at least half the measurements. We show that in this scenario the optimal worst-case estimators should ignore *all* m measurements and be based solely on the a-priori distribution of θ. This result is in sharp contrast with non-adversarial detection theory where even very noisy data can provide some information. This also highlights the power of adversarial manipulation of sensor data since an attacker that has the ability to manipulate only half the sensors, effectively destroys all the information that can be inferred from the full set of sensors.

For the case $n < m/2$, in which the attacker can manipulate less than half the sensors, the optimal estimator typically depends on the sensor data. Moreover, we show that the optimal estimator consists of a threshold rule that compares a Hamming-like distance between the (manipulated) measurement vector and two appropriately defined sets. In general, these sets may be difficult to compute but we provide a procedure to construct the optimal estimator for the boundary case $n = (m-1)/2$, which turns out to be a simple voting scheme. Specific numerical values are provided for the i.i.d. Gaussian case (prior to the adversarial manipulation).

Related Work

Minimax robust detection problems have been extensively studied in the past decades [4]–[6]. A classical approach assumes that the conditional distribution of sensor measurements lies in a set of probability distributions, which is called an uncertainty class. One then identifies a pair of "least

favorable distributions" (LFDs) in the uncertainty class, which conceptually represents the most similar and hardest to distinguish pair of distributions. The robust detector is then designed as a naive-Bayes or Neymann-Pearson detector between the LFDs. While LFDs have been found for a few uncertainty classes, there is no systematic procedure to construct the LFDs and the corresponding estimators, which is the main challenge to apply such approaches in the presence of integrity attacks.

Basar et al. [7], [8] consider the problem of transmitting and decoding Gaussian signals over a communication channel with unknown input from a so-caller "jammer". The unknown input is assumed to be mean square bounded by a constant, which characterizes the capability of the "jammer". Although the mean square bounded assumption is reasonable for analog communications where the attacker is constrained by "energy", it is not practical for cyber attacks on digital communications, where the attacker can change the data arbitrarily as long as the integrity of the sensor is compromised.

The rest of paper is organized as follows: In Section II we formulate the problem of robust detection with n manipulated measurements from m total measurements. In Section III and IV, we consider the optimal detector design for the cases $n \geq m/2$ and $n < m/2$ respectively. Furthermore, in Section V we discuss a special case where $n = (m-1)/2$ and formulate the problem of optimal detector design as an optimization problem. In Section VI we provide a numerical example of i.i.d. Gaussian signals. Finally Section VII concludes the paper.

II. Problem Formulation

The goal is to estimate a binary random variable (r.v.) θ with distribution

$$\theta = \begin{cases} -1 & \text{w.p. } p^- \\ +1 & \text{w.p. } p^+ \end{cases}$$

where $p^-, p^+ \geq 0$ and $p^- + p^+ = 1$. Without loss of generality, we assume that $p^+ \geq p^-$. To estimate θ we have available a vector $y \triangleq [y_1, \ldots, y_m]' \in \mathbb{R}^m$ of m sensor measurements $y_i \in \mathbb{R}$, $i \in \{1, 2, ..., m\}$, each of which is conditionally independent from the others given θ. The conditional probability density (mass) function of each y_i is denoted by

$$P(y_i \in dx|\theta) = F(y_i|\theta)dx.$$

We assume that an attacker wants to increase the probability that we make an error in estimating θ. To this end, the attacker has the ability to manipulate n of the m sensor measurements, but we do not know which n of the m measurements have been manipulated. Formally, this means that our estimate has to rely on a vector $y' \in \mathbb{R}^m$ of *manipulated measurements* defined by

$$y' = y + \gamma \circ u, \tag{1}$$

where $\circ$ is element-wise multiplication and the *sensor-selection* vector γ taking values in

$$\Gamma \triangleq \{\gamma \in \mathbb{R}^m : \gamma_i = 0 \text{ or } 1, \sum_{i=1}^{m} \gamma_i = n\}$$

and the *bias* vector u taking values in $\mathbb{R}^m$. By selecting which values of γ are nonzero, the attacker chooses which of the n sensors will be manipulated. The "level" of manipulation is determined by u.

The estimation problem is formalized as a minimax problem where one wants to select an optimal estimator

$$\hat{\theta} = f(y') = f(y + \gamma \circ u) \tag{2}$$

so as to minimize the probability of error, for a worst case manipulation by the adversary. Following Kerckhoffs' Principle that security should not rely on the obscurity of the system, our goal is to design the estimator $f : \mathbb{R}^m \to \{-1, 1\}$ assuming that f is known to the attacker. We also take the conservative approach that the attacker has full information about the state of the system. Namely, the underlying θ and all the measurements $y_1, \ldots, y_m$ are assumed to be known to the attacker. However, due to limited resources, he can only manipulate n of the m sensors. We assume that the defender knowns how many sensors n can be attacked, but cannot identify them.

To compute the worst-case probability of error that we seek to minimize, we consider given values of θ, y and an estimator f, for which an optimal policy for the attacker can be written as follows:

$$(u, \gamma) = \begin{cases} \underset{u \in \mathbb{R}^m, \gamma \in \Gamma}{\arg\min}\ f(y + \gamma \circ u) & \theta = 1 \\ \underset{u \in \mathbb{R}^m, \gamma \in \Gamma}{\arg\max}\ f(y + \gamma \circ u) & \theta = -1, \end{cases}$$

where the selection of the manipulation pair (u, γ) tries to get the estimate in (2) as low as possible when $\theta = 1$ (ideally as low as -1) or as high as possible when $\theta = -1$ (ideally as high as 1). The min and max are attainable since f only takes binary values.

Under this worst-case attacker policy, a correct decision will be made only when the pair (θ, y) belongs to the set

$$\Big\{(-1, y) : y \in Y^-(f)\Big\} \cup \Big\{(+1, y) : y \in Y^+(f)\Big\} \tag{3}$$

where $Y^+(f)$ and $Y^-(f)$ denotes the set of measurement values $y \in \mathbb{R}^m$ for which the decision of the detector will always be 1 and -1 respectively, regardless of the attacker's action, i.e.,

$$Y^+(f) \triangleq \big\{y \in \mathbb{R}^m : f(y + \gamma \circ u) = 1,\ \forall u \in \mathbb{R}^m, \gamma \in \Gamma\big\},$$
$$Y^-(f) \triangleq \big\{y \in \mathbb{R}^m : f(y + \gamma \circ u) = -1,\ \forall u \in \mathbb{R}^m, \gamma \in \Gamma\big\}.$$

For a given estimator f, the worst-case probability of error $P_e(f)$ is then given by the measure of the set defined in (3) and can be expressed as

$$P_e(f) \triangleq (1 - \beta(f))p^+ + \alpha(f)p^-, \tag{4}$$

with

$$\alpha(f) \triangleq 1 - \sup\{P(y \in S|\theta = -1) : S \in \mathcal{B}(\mathbb{R}^m), S \subseteq Y^-(f)\},$$
$$\beta(f) \triangleq \sup\{P(y \in S|\theta = 1) : S \in \mathcal{B}(\mathbb{R}^m), S \subseteq Y^+(f)\},$$

where $\mathcal{B}(\mathbb{R}^m)$ is the Borel σ-algebra on $\mathbb{R}^m$. One should think of $\alpha(f)$ as the measure of the set $\mathbb{R}^m \setminus Y^-(f)$ conditioned to $\theta = -1$ and of $1 - \beta(f)$ as the measure of the set $\mathbb{R}^m \setminus Y^+(f)$ conditioned to $\theta = +1$. The more complicated definitions of $\alpha(f)$ and $\beta(f)$ use inner measures to make sure that P_e is well defined even if these sets are not measurable.

Formally, the problem under consideration is to determine the optimal estimator f in (2) that minimizes the worst-case probability of error in (4):

$$P_e^* = \inf_f P_e(f).$$

From the discussion above, we can recognize $Y^+(f)$ and $Y^-(f)$ as "good" sets for the estimator, in the sense that when measurements fall in these sets the attacker cannot induce errors. From this perspective, good estimation policies correspond these sets being large. This statement is formalized in the following lemma:

Lemma 1: Given two functions $f, g : \mathbb{R}^m \to \{-1, 1\}$, if $Y^+(g) \supseteq Y^+(f)$ and $Y^-(g) \supseteq Y^-(f)$, then $P_e(g) \leq P_e(f)$.

III. Optimal Detector Design for $n \geq m/2$

In this section we consider the case when half or more of the measurements can be manipulated by the attacker. We show that, in this case, the attacker can render the information provided by the manipulated measurement vector y useless, forcing the optimal estimate to be determined exclusively from the the a-priori distribution of θ.

Theorem 1: If $n \geq m/2$ then the optimal f_* is given by[1]

$$f_*(y) = 1, \quad \forall y \in \mathbb{R}^m,$$

and the corresponding sets Y^+ and Y^- are given by

$$Y^+(f_*) = \mathbb{R}^m, \quad Y^-(f_*) = \emptyset. \qquad \square$$

The following lemma characterizes the relationship between $Y^-(f)$ and $Y^+(f)$ when $n \geq m/2$ and provides a key technical result.

Lemma 2: If $n \geq m/2$, then $Y^-(f) \neq \emptyset$ implies that $Y^+(f) = \emptyset$. $\square$

Proof of Lemma 2. We prove this lemma by contradiction. First it is clear that $m - n \leq n$. Now suppose neither $Y^+(f)$ nor $Y^-(f)$ is empty. As a result, we can find

$$y^+ = [y_1^+, \ldots, y_m^+]' \in Y^+(f),\ y^- = [y_1^-, \ldots, y_m^-]' \in Y^-(f).$$

Now let us consider

$$y = [y_1^+, \ldots, y_n^+, y_{n+1}^-, \ldots, y_m^-]'.$$

Thus,

$$y = y^+ + \gamma_1 \circ (y^- - y^+),$$

where

$$\gamma_1 = [\underbrace{0, \ldots, 0}_{n}, \underbrace{1, \ldots, 1}_{m-n}]'.$$

Therefore, $\gamma_1 \in \Gamma$ and thus by the definition of $Y^+(f)$, $f(y) = 1$. On the other hand,

$$y = y^- + (\mathbf{1} - \gamma_1) \circ (y^+ - y^-),$$

where $\mathbf{1}$ is the one vector. It can be shown that $\mathbf{1} - \gamma_1 \in \Gamma$. Hence, $f(y) = -1$ from the definition of $Y^-(f)$, which contradicts the fact that $f(y) = 1$.

Proof of Theorem 1. By Lemma 2, we know that either $Y^+(f)$ or $Y^-(f)$ must be empty. First suppose that $Y^-(f)$ is empty and hence $\alpha(f) = 1$. As a result

$$P_e(f) = p^+(1 - \beta(f)) + p^-.$$

The minimum is achieved when $Y^+(f) = \mathbb{R}^m$, which implies that $f = 1$ and $P_e(f) = p^-$.
On the other hand, if $Y^+(f)$ is empty, then the optimal $Y^-(f) = \mathbb{R}^m$, $f = -1$ and $P_e(f) = p^+$. Since we assume that $p^+ \geq p^-$, the optimal f is $f_* = 1$ and optimal sets are $Y^+(f_*) = \mathbb{R}^m$ and $Y^-(f_*) = \emptyset$.

IV. Optimal Detector Design for $n < m/2$

We now consider the case when less than half the measurements can be manipulated by the attacker, i.e., $n < m/2$. We show that the optimal estimator is a threshold rule based on a Hamming-like distance between the (manipulated) measurement vector and two appropriately defined sets.

To state the necessary condition for optimality, we need to introduce the following notation: We denote by $d : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^+ \bigcup \{0\}$ the metric induced by the "zero-norm," i.e.,

$$d(x, y) \triangleq \|x - y\|_0,$$

where $\|x\|_0$ is the "zero-norm" of x, which is defined as the number of non-zero entries of the vector x. While the "zero-norm" is not a norm, the function d defined above is a metric. In fact, d can be viewed as an extension of the Hamming distance to continuous-valued vectors. The metric d can be generalized to sets in the usual way: given an element x and two subsets X, Y of $\mathbb{R}^m$, we define

$$d(X, Y) \triangleq \min_{x \in X,\, y \in Y} d(x, y) \quad d(x, Y) \triangleq d(\{x\}, Y). \quad (5)$$

For convenience, we define the distance from any set to the empty set to be infinity: $d(X, \emptyset) = \infty$. The minimum in (5) is always attainable since d takes only integer values.

We also need to introduce a "truncation function": Given an *indexed subset* $\mathcal{I} = \{i_1, i_2, \ldots, i_j\}$ of $\{1, 2, ..., m\}$, we define the function $\mathrm{Trunc}_{\mathcal{I}} : \mathbb{R}^m \to \mathbb{R}^{|\mathcal{I}|}$ by

$$\mathrm{Trunc}_{\mathcal{I}}(y) = \begin{bmatrix} y_{i_1} & y_{i_2} & \cdots & y_{i_j} \end{bmatrix}'.$$

Suppose that for each indexed subset $\mathcal{I} \subset \{1, \ldots, m\}$ of size $m - 2n$ we have a set $S_{\mathcal{I}} \subseteq \mathbb{R}^{m-2n}$. We want to find the largest set $X \subseteq \mathbb{R}^m$ such that $\mathrm{Trunc}_{\mathcal{I}}(X) \subseteq S_{\mathcal{I}}$ for each

[1] Recall that we are assuming that $p^+ \geq p^-$.

$\mathcal{I}$ respectively. It is easy to see that X can be defined in the following way:

$$X \triangleq \{y \in \mathbb{R}^m : \mathrm{Trunc}_{\mathcal{I}}(y) \in S_{\mathcal{I}}, \forall |\mathcal{I}| = m - 2n\}. \quad (6)$$

We define the class of such set X parameterized by $S_{\mathcal{I}}$s as $\mathcal{X}_{m,n}$.

Definition 1: Two sets $X_1, X_2 \in \mathcal{X}_{m,n}$ are called mutually exclusive if and only if

$$X_1 \triangleq \{y \in \mathbb{R}^m : \mathrm{Trunc}_{\mathcal{I}}(y) \in S_{\mathcal{I}}, \forall |\mathcal{I}| = m - 2n\},$$
$$X_2 \triangleq \{y \in \mathbb{R}^m : \mathrm{Trunc}_{\mathcal{I}}(y) \in \mathbb{R}^{m-2n} \backslash S_{\mathcal{I}}, \forall |\mathcal{I}| = m - 2n\},$$

for some $S_{\mathcal{I}}$s.

Theorem 2: The optimal estimator f_* is of the form

$$f^*(y) = \begin{cases} 1 & d(y, X^-) \geq d(y, X^+) \\ -1 & d(y, X^-) < d(y, X^+), \end{cases} \quad (7)$$

where $X^+, X^- \in \mathcal{X}_{m,n}$ are mutually exclusive. Moreover, $X^+ \subseteq Y^+(f^*)$ and $X^- \subseteq Y^-(f^*)$. □

A. Proof of Theorem 2

The remainder of this section is mostly devoted to the proof of Theorem 2, which requires several intermediate results.

For a given set $\mathcal{I}$ and estimator f, in the sequel we denote by $Y^-_{\mathcal{I}}(f)$ and $Y^+_{\mathcal{I}}(f)$ the image of $Y^-(f)$ and $Y^+(f)$, respectively, under the function $\mathrm{Trunc}_{\mathcal{I}}$. As stated in the following result, it turns out that these sets are always disjoint:

Lemma 3: if $n < m/2$, for every estimator f and index subset $\mathcal{I}$ of size $|\mathcal{I}| = m - 2n$, we have

$$Y^-_{\mathcal{I}}(f) \bigcap Y^+_{\mathcal{I}}(f) = \emptyset.$$

Proof of Lemma 3. We prove the statement by contradiction. Without loss of generality, we assume that $\mathcal{I} = \{1, \ldots, m - 2n\}$, and

$$Y^-_{\mathcal{I}} \bigcap Y^+_{\mathcal{I}} \neq \emptyset.$$

As a result, there exist

$$y^+ = [y_1, \ldots, y_{m-2n}, y^+_{m-2n+1}, \ldots, y^+_m]' \in Y^+(f),$$

and

$$y^- = [y_1, \ldots, y_{m-2n}, y^-_{m-2n+1}, \ldots, y^-_m]' \in Y^-(f).$$

Now let us consider

$$y = [y_1, \ldots, y_{m-2n}, y^+_{m-2n+1} \cdots, y^+_{m-n}, y^-_{m-n+1}, \ldots, y^-_m]'.$$

It can be easily seen that there are n elements in y that differ from $y^+ \in Y^+(f)$ and another n elements in y that differ from $y^- \in Y^-(f)$. As a result, $f(y) = 1$ from the definition of $Y^+(f)$ and $f(y) = -1$ from the definition of $Y^-(f)$, which is an absurd.

For every estimator f, it is easy to see that

$$Y^-(f) \subseteq \mathcal{Y}^-(f)$$
$$\triangleq \{y \in \mathbb{R}^m : \mathrm{Trunc}_{\mathcal{I}}(y) \in Y^-_{\mathcal{I}}(f), \forall |\mathcal{I}| = m - 2n\}.$$

From Lemma 3, we can conclude that $Y^+_{\mathcal{I}}(f) \subseteq R^{2m-n} \backslash Y^-_{\mathcal{I}}(f)$ and therefore, $Y^+(f)$ is upper bounded by

$$Y^+(f) \subseteq \mathcal{Y}^+(f)$$
$$\triangleq \{y \in \mathbb{R}^m : \mathrm{Trunc}_{\mathcal{I}}(y) \in \mathbb{R}^{m-2n} \backslash Y^-_{\mathcal{I}}(f), \forall |\mathcal{I}| = m - 2n\}.$$

Therefore $\mathcal{Y}^-(f), \mathcal{Y}^+(f) \in \mathcal{X}_{m,n}$ and are mutually exclusive. We will prove that there exists a function g of the form (7), for which $\mathcal{Y}^-(f) \subseteq Y^-(g)$ and $\mathcal{Y}^+(f) \subseteq Y^+(g)$. Before that, we want to provide an inequality on the distance between an arbitrary vector y and the sets $\mathcal{Y}^+(f)$ and $\mathcal{Y}^-(f)$, the proof of which is omitted due to space limit.

Lemma 4: $d(y, \mathcal{Y}^-(f)) + d(y, \mathcal{Y}^+(f)) \geq 2n + 1$. □

We can now prove the main technical result needed for the proof of Theorem 2:

Lemma 5: Consider the function $g : \mathbb{R}^m \to \{-1, 1\}$, defined as

$$g(y) = \begin{cases} 1 & d(y, \mathcal{Y}^-(f)) \geq d(y, \mathcal{Y}^+(f)) \\ -1 & d(y, \mathcal{Y}^-(f)) < d(y, \mathcal{Y}^+(f)). \end{cases} \quad (8)$$

Then

$$Y^-(f) \subseteq \mathcal{Y}^-(f) \subseteq Y^-(g),\ Y^+(f) \subseteq \mathcal{Y}^+(f) \subseteq Y^+(g).$$

□

Proof of Lemma 5. We first prove that

$$\mathcal{Y}^-(f) \subseteq Y^-(g).$$

Consider an arbitrary $y \in \mathcal{Y}^-(f)$. We need to prove that for any $u \in \mathbb{R}^m$ and $\gamma \in \Gamma$, $g(y + \gamma \circ u) = -1$. From definition,

$$d(y + \gamma \circ u, \mathcal{Y}^-(f)) \leq d(y + \gamma \circ u, y) \leq n.$$

By Lemma 4,

$$d(y + \gamma \circ u, \mathcal{Y}^+(f)) + d(y + \gamma \circ u, \mathcal{Y}^-(f)) \geq 2n + 1,$$

which implies that

$$d(y + \gamma \circ u, \mathcal{Y}^+(f)) \geq n + 1.$$

As a result,

$$d(y + \gamma \circ u, \mathcal{Y}^-(f)) < d(y + \gamma \circ u, \mathcal{Y}^+(f)).$$

Therefore $g(y + \gamma \circ u) = -1$, which implies that $\mathcal{Y}^-(f) \subseteq Y^-(g)$. Similarly, one can prove that $\mathcal{Y}^+(f) \subseteq Y^+(g)$.

Proof of Theorem 2. Theorem 2 now follows from Lemma 1, since $Y^-(f) \subseteq Y^-(g)$ and $Y^+(f) \subseteq Y^+(g)$ together imply that $P_e(f) \geq P_e(g)$.

Combining Lemma 5 and Theorem 2, we have the following corollary.

Corollary 1: The optimal X^+ and X^- of f^* are the solutions of the following optimization problem[2]:

$$\begin{aligned} \underset{X^+, X^-}{\text{minimize}} \quad & 1 - P(y \in X^+ | \theta = 1) - P(y \in X^- | \theta = -1) \\ \text{subject to} \quad & X^+,\ X^- \in \mathcal{X}_{m,n}, \\ & X^+,\ X^- \text{ are mutually exclusive.} \end{aligned}$$

[2]Inner measure should be used if X^+ or X^- are not measurable.

The key challenge in applying Theorem 2 and Corollary 1 is that they do not provide a construction for the set X^+, X^- that lead to the optimal f^*, potentially requiring one to search for the optimal estimator by ranging over all possible sets in $\mathcal{X}_{m,n}$. However, we shall see in Section V that we can use this general result to find the optimal estimator, at least for the case $m = 2n + 1$.

V. Optimal Detector for $n = (m-1)/2$

In this section, we construct the optimal detector for the case where $n = (m-1)/2$. From Theorem 2, we know that the optimal estimator can be constructed by choosing an 'appropriate" family of sets $S_{\mathcal{I}}$. It turns out that when $n = (m-1)/2$ this family of sets has a particularly simple structure:

Theorem 3: If $m - 2n = 1$, the family of sets $S_{\{i\}}$ that give the optimal estimator f^* in (7) is of the form

$$S_{\{i\}} = T_i(\eta_i), \quad \forall i \in \{1, 2, ..., m\} \tag{9}$$

where each $\eta_i \in \mathbb{R} \cup \{-\infty, +\infty\}$,

$$T_i(\eta) \triangleq \left\{ y_i \in \mathbb{R} : \log\left(\frac{F(y_i|\theta_i = 1)}{F(y_i|\theta_i = -1)}\right) < \eta \right\}.$$

By convention, $T_i(\infty) = \mathbb{R}$ and $T_i(-\infty) = \emptyset$.

Before proving Theorem 3, we note that one can implement the optimal estimator in (7) without actually computing $d(y, X^-)$ and $d(y, X^+)$. When X^+ or X^- are empty, then one of the distances in (7) is $+\infty$ and f^* is simply a constant. When none of these sets is empty, it is straightforward to show that

$$d(y, X^-) = \left|\{i : y_i \in T_i(\eta_i)\right|, \quad d(y, X^+) \triangleq \left|\{i : y_i \notin T_i(\eta_i)\right|$$

and the detection algorithm can be implemented as the following voting process:

- The detector computes m individual estimates $\hat{\theta}_i$ by a Neymann-Pearson detector based on individual (possibly manipulated) measurements y_i':

$$\hat{\theta}_i \triangleq \begin{cases} -1 & y_i' \in T_i(\eta_i) \\ 1 & y_i' \notin T_i(\eta_i) \end{cases}$$

- The optimal estimate $\hat{\theta}$ is obtained by voting:

$$\hat{\theta} = \begin{cases} -1 & \text{at least } n+1 \text{ estimates } \hat{\theta}_i = -1 \\ +1 & \text{less than } n+1 \text{ estimates } \hat{\theta}_i = -1 \end{cases}$$

A. Proof of Theorem 3

We start by noting that when $m - 2n = 1$ the sets X^- and X^+ are especially simple to compute:

$$X^- = \left\{ y \in \mathbb{R}^m : \text{Trunc}_i(y) \in S_{\{i\}}, \forall i = 1, \ldots, m \right\} = \prod_{i=1}^m S_{\{i\}}$$

$$X^+ = \left\{ y \in \mathbb{R}^m : \text{Trunc}_i(y) \in \mathbb{R}\backslash S_{\{i\}}, \forall i = 1, \ldots, m \right\} = \prod_{i=1}^m \mathbb{R}\backslash S_{\{i\}},$$

where $\prod_i S_{\{i\}}$ should be interpreted as a Cartesian product. The following result is a straightforward consequence of the fact that X^- and X^+ can be written as Cartesian products:

Lemma 6: If $X^+ \neq \emptyset$ and $X^- \neq \emptyset$, then $\text{Trunc}_i(X^-) = S_{\{i\}}$ and $\text{Trunc}_i(X^+) = \mathbb{R}\backslash S_{\{i\}}$.

The following result essentially states that the inner measure of Cartesian products is the product of inner measure of each set. The detail of the proof is omitted due to space limit.

Lemma 7: Let

$$\alpha_i = 1 - \sup\{P(y_i \in S|\theta = -1) : S \in \mathcal{B}(\mathbb{R}), S \subseteq S_{\{i\}}\},$$
$$\beta_i = \sup\{P(y_i \in S|\theta = 1) : S \in \mathcal{B}(\mathbb{R}), S \subseteq \mathbb{R}\backslash S_{\{i\}}\}.$$

If X^-, $X^+ \neq \emptyset$, then the following holds:

$$\alpha = 1 - \prod_{i=1}^m (1 - \alpha_i), \; \beta = \prod_{i=1}^m \beta_i.$$

We are now ready to prove Theorem 3 by leveraging the independence of y_i.

Proof of Theorem 3. It is easy to see that if X^+ (X^-) is empty, then $f = -1$ ($f = 1$), which implies that $S_{\{i\}} = T_i(\infty)$ ($S_{\{i\}} = T_i(-\infty)$). Now assume that X^+ and X^- are not empty, by Lemma 7,

$$P_e(f) = 1 - p^+ \prod_{i=1}^m \beta_i - p^- \prod_{i=1}^m (1 - \alpha_i).$$

Suppose the optimal α_i, β_i are α_i^*, β_i^*. As a result, we know that

$$\begin{aligned} P_e^* &= 1 - \left[p^+ \prod_{j \neq i} \beta_j^*\right] \beta_i - \left[p^- \prod_{j \neq i} (1 - \alpha_j^*)\right] (1 - \alpha_i) \\ &= a_i^* \alpha_i - b_i^* \beta_i + c_i^*, \end{aligned}$$

where

$$a_i^* = \left[p^- \prod_{j \neq i} (1 - \alpha_j^*)\right], b_i^* = \left[p^+ \prod_{j \neq i} \beta_j^*\right], \; c_i^* = 1 - a_i^*.$$

By the Bayes Risk Criterion [9], the optimal $S_{\{i\}}$ must be of form (9), with $\eta_i = log(a_i^*/b_i^*)$.

Remark 1: It can be shown that β_i is a function of α_i (which corresponds to the ROC curves of the Neymann-Pearson detector of y_i), when $S_{\{i\}}$ is of the form (9). Due to Corollary 1, we know that

$$P_e^* = \min_{\alpha_i} 1 - p^+ \prod_{i=1}^m \beta_i - p^- \prod_{i=1}^m (1 - \alpha_i),$$

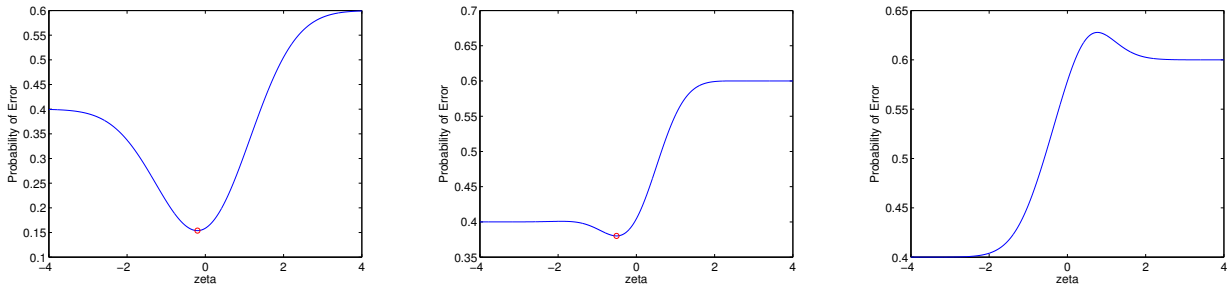
which can be solved numerically. □

VI. I.I.D. Gaussian Case

We now specialize our results for *i.i.d. Gaussian measurement* y_i. In particular, we assume that

$$y_i = a\theta + v_i,$$

where $a > 0$ is constant and v_is denote i.i.d. Gaussian variables. Without loss of generality, we assume that these



(a) $m = 1,\ n = 0$ (b) $m = 3,\ n = 1$ (c) $m = 5,\ n = 2$

Fig. 1. Probability of Error v.s. Threshold ζ

variables have zero mean and unit variance. It is easy to prove that $S_{\{i\}}$ in Theorem 3 are of the form

$$S_{\{i\}} = T(\eta_i) = \{y_i \in \mathbb{R} : y_i < \zeta_i\}, \tag{10}$$

with $\zeta_i = \eta_i/2a$. Moreover, the following results uses symmetry to provide an even tighter characterization of the sets corresponding to the optimal estimator.

Theorem 4: In the case of i.i.d. Gaussian measurements and $m - 2n = 1$, the optimal worst-case probability of error is given by

$$P_e^* = 1 - \sup_{\zeta}\ p^+ \left[Q(\zeta - a)\right]^m + p^- \left[Q(-\zeta - a)\right]^m, \tag{11}$$

where

$$Q(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du.$$

Moreover, the S_is of the optimal estimator f^* in (7) are symmetric and of the form

$$S_{\{i\}} = \{y_i \in \mathbb{R} : y_i < \zeta\}, \quad \forall i \in \{1, 2, ..., m\} \tag{12}$$

for any $\zeta \in \mathbb{R} \cup \{-\infty, +\infty\}$ that achieves the supremum in (11). □

Proof. The proof uses the fact that $Q(x)$ is a logarithmically concave function. The detailed proof is omitted due to space limit.

Remark 2: The main difference between Theorem 4 and 3 is that all the individual thresholds in Theorem 4 are essentially the same, which reduces the search space further from $\mathbb{R}^m$ to $\mathbb{R}$.

In Figure 1 we plot the probability of error versus the threshold ζ for different pairs of m, n. The parameters are chosen as follows:

$$p^+ = 0.6,\ p^- = 0.4,\ a = 1.$$

The optimum for $m = 1,\ n = 0$ is $\zeta = -0.202,\ P_e = 0.154$. The optimum for $m = 3,\ n = 1$ is $\zeta = -0.508,\ P_e = 0.380$. For the case $m = 5,\ n = 2$, the optimal ζ is actually $-\infty$. Therefore, the optimal detector is simply $f_* = 1$.

VII. Conclusion

In this paper we consider the problem of designing detectors able to minimize the probability of error in the face of n corrupted measurements due to integrity attacks on a subset of the sensor pool. The problem is posed as a minimax optimization where the goal is to design the optimal detector against all possible attacker's strategies. We show that if the attacker can manipulate at least half of the m measurements ($n \geq m/2$) then the optimal worst-case estimator should ignore *all* m measurements and be based solely on the a-priori information. When the attacker can manipulate less than of half of the measurements ($n < m/2$), we show that the optimal estimator is a threshold rule based on a Hamming-like distance between the manipulated measurement vector and two appropriately defined sets. For a particular case ($m = 2n+1$) we were able to compute the optimal detector, showing that it consists of a simple voting scheme. We further apply the results to $n = (m-1)/2$ case and i.i.d. Gaussian case.

References

[1] T. M. Chen, "Stuxnet, the real start of cyber warfare? [editor's note]," *IEEE Network*, vol. 24, no. 6, pp. 2–3, 2010.

[2] D. P. Fidler, "Was stuxnet an act of war? decoding a cyberattack," *IEEE Security & Privacy*, vol. 9, no. 4, pp. 56–59, 2011.

[3] A. A. Cárdenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems," in *HOTSEC'08: Proceedings of the 3rd conference on Hot topics in security*. Berkeley, CA, USA: USENIX Association, 2008, pp. 1–6.

[4] P. J. Huber, "A robust version of the probability ratio test," *The Annals of Mathematical Statistics*, vol. 36, no. 6, pp. pp. 1753–1758, 1965. [Online]. Available: http://www.jstor.org/stable/2239116

[5] P. J. Huber and V. Strassen, "Minimax tests and the Neyman-Pearson lemma for capacities," *The Annals of Statistics*, vol. 1, no. 2, pp. pp. 251–263, 1973. [Online]. Available: http://www.jstor.org/stable/2958011

[6] S. A. Kassam and H. V. Poor, "Robust techniques for signal processing: A survey," *Proceedings of the IEEE*, vol. 73, no. 3, pp. 433–481, 1985.

[7] T. Basar and Y. W. Wu, "Solutions to a class of minimax decision problems arising in communication systems," *Journal of Optimization Theory and Applications*, vol. 51, pp. 375–404, 1986.

[8] R. Bansal and T. Basar, "Communication games with partially soft power constraints," *Journal of Optimization Theory and Applications*, vol. 61, pp. 329–346, 1989.

[9] L. Scharf, *Statistical Signal Processing*. Prentice Hall, 1990.