Detection in Adversarial Environments

K. G. Vamvoudakis[1], Member, IEEE J. P. Hespanha[1], Fellow, IEEE B. Sinopoli[2], Member, IEEE Y. Mo[2]

Abstract

We propose new game theoretic approaches to estimate a binary random variable based on sensor measurements that may have been corrupted by a cyber-attacker. The estimation problem is formulated as a zero-sum partial information game in which a detector attempts to minimize the probability of an estimation error and an attacker attempts to maximize this probability. While this problem can be solved exactly by reducing it to the computation of the value of a matrix, this approach is computationally feasible only for a small number of sensors. The two key results of this paper provide complementary computationally efficient solutions to the construction of the optimal detector. The first result provides an explicit formula for the optimal detector but it is only valid when the number of sensors is roughly smaller than two over the probability of sensor errors. In contrast, the detector provided by the second result is valid for an arbitrary number of sensor. While it may result in a probability of estimation error that is ϵ above the minimum achievable, we show that this error ϵ is small when the number of sensors is large, which is precisely the case for which the first result does not apply.

Index Terms

Adversarial detection, byzantine sensors, cyber security, zero-sum games, estimation.

I. Introduction

Embedded sensors, computation, and communication have enabled the development of sophisticated sensing devices [17] for a wide range of cyber physical applications that include safety

*This material is based upon work supported in part by ARO MURI Grant number W911NF0910553.

[1]K. G. Vamvoudakis, and J. P. Hespanha are with the Center for Control, Dynamical-systems and Computation (CCDC), University of California, Santa Barbara, CA 93106-9560 USA e-mail: kyriakos@ece.ucsb.edu.

[2]B. Sinopoli, and Y. Mo are with the Electrical and Computer Engineering Department, Carnegie Mellon University, Pittsburgh, PA, USA.

monitoring, health-care, surveillance, traffic monitoring, and military applications [5, 6, 15, 20]. However, the deployment of such devices has been slowed by concerns regarding their vulnerability to both stochastic failures and cyber attacks. Of particular concern are scenarios in which an attacker *gains access to the computing platform of a sensing device and manipulates the output reported by the device to compromise any decision based on that data*. This scenario forces system designers to re-think basic estimation problems in light of security concerns.

In traditional estimation problems one attempts to determine the value of a physical variable that cannot be measured directly based on a set of "noisy" measurements of that variable [11]. Typically, some form of probabilistic structure is assumed to model how the measurements relate to the true value of the variable to be estimated. This type of framework is adequate, when the measurements fluctuate around the variable's true value, e.g. due to microscopic thermal fluctuations. However, things are quite different when the measurement device can be controlled by an entity that actively attempts to degrade the estimation process.

The most basic mechanism to overcome stochastic measurements errors relies on the use of redundancy. When multiple sensors provide redundant and independent measurements about a variable that needs to be estimated, the confidence on the estimate increases with the number of sensors. When some of these sensors are being controlled by an adversary that wants to maximize the estimation error, the independence assumption is generally not valid and the magnitude of an estimation error scales differently with the number of sensors. The goal of this paper is to provide insights regarding what happens in such situations.

We consider the problem of estimating the value of a binary random variable based on measurements provided by a group of binary sensors. We assume that such measurements incorporate two types of errors: purely stochastic errors that are responsible for bit-flips with a given probability, and adversarial errors that are controlled by an adversary that has infiltrated a subset of the sensors. Which sensors have been manipulated is not known a-priori to the detection system. We shall see that the (optimal) adversarial errors may actually be stochastic (corresponding to mixed policies), with probability distributions carefully selected by the attacker to maximize the probability of an estimation error. In general, these distributions will be a function of the value of the variable to be estimated. A key novelty of the work presented here with respect to classical problems of Byzantine faults [10] is that we do not assume perfect

sensors, i.e., even the sensors that have not been manipulated can report an incorrect value with a given probability $p_{\text{error}} > 0$, typically due to limitations of the physical sensing mechanism.

The adversarial estimation problem described above is formulated as a zero-sum game between a player that wants to estimate the binary random variable with minimal probability of an estimation error, henceforth called the *detector*, and a player that wants to maximize the same probability of error, henceforth called the *attacker*. This is a game of partial information [3, 8, 12] in that the decision maker only has access to "noisy" sensor measurements that have been corrupted both by stochastic and by adversarial errors and does not know which sensor measurements have been compromised by the attacker. Similarly, the attacker also only has partial information since, while she may know the true value of the variable to be estimated, she does not know the values of the measurements that are being reported by the sensors that she has not infiltrated.

To model the fact that the detector may not be certain whether or not an attacker may actually have infiltrated some of the sensors, we introduce a "probability of attack" parameter p_{attack} that reflects how certain the detector is about the existence of a malicious attacker. An interesting feature of the solutions obtained is that the optimal estimation policies are largely insensitive to this parameter. This is convenient because p_{attack} would typically be hard to guess.

The adversarial estimation problem considered here can be reduced to the computation of the (mixed) saddle-point of a zero-sum matrix game (cf., Section II). However, even for a relatively small number of sensors n, the matrix can become very large. To overcome this difficulty we provide two complementary approaches that scale to a very large number of sensors. The main result of Section III (Theorem 1) provides an explicit formula for the optimal estimator (i.e., the saddle-point policy for the detector) and the corresponding probability of estimation error (i.e., value of the game) that is valid when the number of sensors is roughly below two over the probability p_{error} of sensor error. Somewhat unexpectedly the optimal estimator is a mixed policy that randomizes between a majority rule (i.e., pick the value reported by most sensors) and another rule that can go against the majority.

The main result of Section IV (Theorem 3) provides an explicit formula for a detection policy that is valid for an arbitrary number of sensors. While this detection policy is not necessarily a saddle point, we show that it leads to a probability of estimation error that is, at most, ϵ

larger than the optimal. Moreover, the value of ϵ decays to zero with $1/\sqrt{n}$, which means that the estimator provided in Section IV is very close to optimal precisely when the estimator in Section III does not apply.

The results presented in Section III and IV are obtained using different approaches. In Section III, the large zero-sum matrix game is reduced to a 2×2 matrix game through successive applications of policy domination and the result follows from the direct solution of the 2×2 matrix games. In Section IV, we approximate the original game, where the players make discrete choices, with a continuous game where the choices are continuous. This relaxation leads to a game for which we can find the (exact) saddle point, which turns out to be an ϵ-saddle-point for the original (discrete) game. The solution of the continuous game (Theorem 2) may be of independent interest for other adversarial estimation problems.

Related Work

There is a large body of literature regarding game theoretical approaches to cyber and network security [1–3, 16]. Byzantine attacks have their root in the Byzantine generals problem where the traitor generals want to prevent the loyal generals from reaching an agreement [10]. Sensor fusion with Byzantine sensors is presented in [9], where the authors use random binning in sensor polling to force a Byzantine sensor to either act honestly or reveal its Byzantine identity. This random binning is not needed when more than half of the sensors are honest. The authors in [7], describe the Byzantine problem as a zero-sum game problem in which the attacker's policy is to manipulate the measurements of the sensor network and the defender measures the sensor value before picking an action, but without providing any closed-form policies. The game is solved as a pair of dual linear programming problems and it is shown that deception becomes more difficult when sensor redundancy is used. The authors in [4] propose a game theoretical model for virtual coordinate systems that allows hosts in the Internet to determine latency to arbitrary hosts based on information provided by a subset of sensors. The Byzantine adversary knows how and what defense strategies are used and adjusts his strategies accordingly. The defender, on the other side, uses an adaptive threshold to decide if the data should be accepted by the system or not and thus deter adaptive adversaries. Game theoretic solutions for sensor networks based on cooperation and selfishness have been reported in [14, 18], where each node needs to decide whether to forward or not a measurement based on appropriate payoff functions.

Our recent work in [13] also deals with estimation under stochastic errors and cyber attacks. In that work we considered a full information game for the attacker and provided minimax pure policies that may not be saddle-point policies, which leads to more conservative solutions for the detector. A subset of the results in Section III appeared in the conference paper [19].

II. Problem formulation

The goal of this paper is to estimate the value of a binary random variable X with Bernoulli distribution

$$\mathrm{P}(X=1) = 1 - \mathrm{P}(X=0) = p \in (0,1), \tag{1}$$

based on a vector $Y \coloneqq (Y_1, Y_2, \ldots, Y_n)$ of n binary "noisy" sensor measurements, where the measurements Y_i, $i \in \{1, 2, \ldots, n\}$ are assumed conditionally independent, given X. Specifically,

$$\mathrm{P}(Y_i = 1 \,|\, X, Y_{j\neq i}) = \begin{cases} p_{\mathrm{err}} & X = 0, \\ 1 - p_{\mathrm{err}} & X = 1, \end{cases} \tag{2}$$

where $p_{\mathrm{err}} \in [0,1]$ denotes the sensor error probability. Setting us apart from standard estimation problems, we consider a scenario where an estimate $\hat{X}$ of X needs to be constructed based on version $Z \coloneqq (Z_1, Z_2, \ldots, Z_n)$ of the measurement vector Y that may have been "corrupted" by an attacker. It is assumed that, with a given probability $p_{\mathrm{attack}} \in [0,1]$, the attacker manipulated the readings of $m \leqslant n$ sensors and therefore only $m - n$ of the Z_i match the corresponding Y_i, but the estimator of X does not know which. The probability $p_{\mathrm{attack}} \in [0,1]$ should be viewed as a design parameter that reflects how certain the estimator is that the measurements have been manipulated. For $p_{\mathrm{attack}} = 0$, we recover a standard estimation problem with purely stochastic measurement errors.

The problem under consideration can be viewed as a two-player partial information game: The *detector* must select its estimate $\hat{X}$ based solely on the vector Z of possibly corrupted sensor readings. Because the detector does not know which sensors have been manipulated, its decision must be solely based on the total number of zeros and ones in the vector Z, which corresponds to the selection of the *estimation policy* $\mu : \{0, 1, \ldots, n\} \to \{0, 1\}$ that is used to compute the estimate

$$\hat{X} = \mu\Big(\sum_{i=1}^{n} Z_i\Big). \tag{3}$$

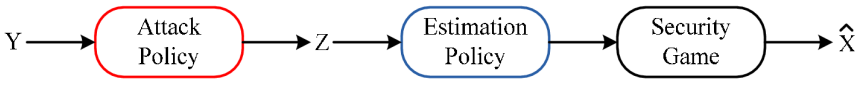

Fig. 1. Detection-Attack Model

Since the domain of μ has $n+1$ elements and its codomain has 2 elements, the set $\mathcal{U}$ of all possible estimation policies contains 2^{n+1} policies.

We assume that the *attacker* knows the true value of X and bases her decision on how to corrupt the m measurements of the manipulated sensors as a function of X. Since the attacker is not assumed to know the values reported by the remaining sensors, she also suffers from partial information. We can thus view the *attack policy* as a function $\delta : \{0,1\} \to \{0,1,\ldots,m\}$, with the understanding that $\delta(X)$ determines how many of the m sensors that have been manipulated will report a zero (the other $m-\delta(X)$ will report a one). Since the domain of δ has 2 elements and its codomain has $m+1$ elements, the set $\mathcal{D}$ of all possible attack policies contains $(m+1)^2$ policies.

The model just described is illustrated in Figure 1 and allow us to define adversarial estimation as a zero-sum game in which the detector selects a policy $\mu \in \mathcal{U}$ and the attacker a policy $\delta \in \mathcal{D}$ so to minimize and maximize, respectively, the probability of an estimation error

$$\mathrm{P}_{\mu,\delta}(\hat{X} \neq X), \tag{4}$$

where the subscript $_{\mu,\delta}$ in the probability measure emphasizes the fact that the probability of an estimation error depends on the players' policies. Since the sets of policies are finite, we have a (finite) matrix game defined by a 2^{n+1} by $(m+1)^2$ matrix

$$A \coloneqq \left[a_{ij}\right]_{2^{n+1}\times(m+1)^2},$$

where a_{ij} denotes the probability of an estimation error (4) corresponding to the ith estimation policy in $\mathcal{U}$ and the jth attack policy in $\mathcal{D}$. In general, this game does not have pure saddle-point equilibria so the players will seek for mixed policies, which correspond to selecting probability distributions over the sets of actions $\mathcal{U}$ and $\mathcal{D}$.

The following result, proved in the appendix, can be used to compute the matrix (4).

Lemma 1: When the detector utilizes a policy (3) and the attacker utilizes a policy δ that sets to 0 and to 1 a number of sensors equal to $\delta(X)$ and $m-\delta(X)$, respectively, the probability of an estimation error is given by

$$\begin{aligned}\mathrm{P}_{\mu,\delta}(\hat{X}\neq X) = (1-p)\Bigg(&p_{\text{attack}}\sum_{k=m-\delta(0)}^{n-\delta(0)}\mu(k)\binom{n-m}{k-m+\delta(0)}p_{\text{err}}^{k-m+\delta(0)}(1-p_{\text{err}})^{n-k-\delta(0)}\\&+(1-p_{\text{attack}})\sum_{k=0}^{n}\mu(k)\binom{n}{k}p_{\text{err}}^{k}(1-p_{\text{err}})^{n-k}\Bigg)\\+p\Bigg(&p_{\text{attack}}\sum_{k=m-\delta(1)}^{n-\delta(1)}\big(1-\mu(k)\big)\binom{n-m}{k-m+\delta(1)}(1-p_{\text{err}})^{k-m+\delta(1)}p_{\text{err}}^{n-k-\delta(1)}\\&+(1-p_{\text{attack}})\sum_{k=0}^{n}\big(1-\mu(k)\big)\binom{n}{k}(1-p_{\text{err}})^{k}p_{\text{err}}^{n-k}\Bigg).\end{aligned}\tag{5}$$

□

III. Closed-form solution for exact saddle-point

We have seen that the estimation problem can be seen as a matrix game with 2^{n+1} policies for the detector and $(m+1)^2$ policies for the attacker. It turns out that the exponential complexity in the number of sensors n can be removed using policy domination. For simplicity of presentation, we show this for the case where it is equally likely that X is 0 or 1 and the number of sensors is odd (allowing for tie-breaking). When $p=1/2$ in (1), we have perfect symmetry between the cases $X=0$ and $X=1$, which means that both players should treat 0 and 1 similarly. In particular, if the detector uses the estimate $\hat{X}=1$ when the vector Z has k 1's, then it should use the estimate $\hat{X}=0$ when the vector Z has k 0's and therefore we can restrict our attention to estimation policies for which

$$\mu(k) = 1-\mu(n-k).\tag{6}$$

Similarly, if the attacker decides to set to 0 a certain number of sensors when $X=0$, then it should set to 1 the same number of sensors when $X=1$ and therefore we can restrict our attention to attack policies for which

$$\delta(0) = m-\delta(1).\tag{7}$$

In this case, we can provide explicit formulas for mixed saddle-point policies and for the value of the game. This result is formulated in terms of the following pure policies:

1) We define the detector's *majority rule* to be the pure policy

$$\mu\Big(\sum_{i=1}^{n} Z_i\Big) = \mu_{\text{majority}}\Big(\sum_{i=1}^{n} Z_i\Big) \coloneqq \begin{cases} 0 & \sum_{i=1}^{n} Z_i \leqslant \frac{n-1}{2} \\ 1 & \sum_{i=1}^{n} Z_i \geqslant \frac{n+1}{2}, \end{cases}$$

which corresponds to setting $\hat{X} = 0$ if more than half the sensors reported the value 0.

2) We define the detector's *no-consensus rule* to be the pure policy

$$\mu\Big(\sum_{i=1}^{n} Z_i\Big) = \mu_{\text{no-consensus}}\Big(\sum_{i=1}^{n} Z_i\Big) \coloneqq \begin{cases} 0 & 0 < \sum_{i=1}^{n} Z_i \leqslant \frac{n-1}{2} \text{ or } \sum_{i=1}^{n} Z_i = n \\ 1 & n > \sum_{i=1}^{n} Z_i \geqslant \frac{n+1}{2} \text{ or } \sum_{i=1}^{n} Z_i = 0. \end{cases}$$

This somewhat unexpected policy is like the majority rule, except that if all sensors agree on a particular value (i.e., $Z_i = 1$, $\forall i$ or $Z_i = 0$, $\forall i$), the estimate $\hat{X}$ should take the opposite value.

3) We define the attacker's *deception rule* to be the pure policy

$$\delta(X) = \delta_{\text{deception}}(X) \coloneqq \begin{cases} 0 & X = 0 \\ m & X = 1 \end{cases}$$

that, when $X = 0$ sets all m manipulated sensors equal to 1 and when $X = 1$ sets all m sensors equal to 0.

4) We define the attacker's *no-deception rule* to be the pure policy

$$\delta(X) = \delta_{\text{no-deception}}(X) \coloneqq \begin{cases} m & X = 0 \\ 0 & X = 1 \end{cases}$$

that, when $X = 0$ sets all m manipulated sensors equal to 0 and when $X = 1$ sets all m sensors equal to 1.

Theorem 1: Consider $p = 1/2$ in (5) and an odd number of sensors $n > 2$, for which

$$m \leqslant \min\left\{\frac{n-1}{2}, \frac{n+1}{2} - \frac{p_{\text{err}}}{1-p_{\text{err}}}\frac{n-1}{2}\right\} \tag{8}$$

with n and p_{err} sufficiently small so that,

$$p_{\text{err}} \leqslant \frac{2}{n+1}, \tag{9}$$

$$p_{\text{err}} \leqslant \left(1 + \frac{\frac{n-1}{2}!(n-m)!}{\frac{n-2m+1}{2}!}\right)^{-1}, \tag{10}$$

and, for the case $m \geqslant 2$, one also needs

$$p_{\text{attack}} \leqslant \frac{1}{1 + \frac{1}{n}\binom{n-m}{m-1}\frac{p_{\text{err}}^{n-2m+1}(1-p_{\text{err}})^{m-1}}{p_{\text{err}}(1-p_{\text{err}})^{n-1} - p_{\text{err}}^{n-1}(1-p_{\text{err}})}}. \tag{11}$$

In this case, the value of the game is given by

$$v^* = \alpha + p_{\text{attack}} \min\left\{\gamma, \frac{\gamma(1-p_{\text{err}})^{n-m} + \beta\gamma + \rho(p_{\text{err}}^{n-m} - \beta)}{(1-p_{\text{err}})^{n-m} + p_{\text{err}}^{n-m}}\right\} \tag{12}$$

and a mixed saddle-point policy corresponds to selecting

$$\begin{cases} \mu_{\text{majority}} & \text{w.p. } 1 - y_2 \\ \mu_{\text{no-consensus}} & \text{w.p. } y_2, \end{cases} \tag{13}$$

$$\begin{cases} \delta_{\text{deception}} & \text{w.p. } 1 - z_2 \\ \delta_{\text{no-deception}} & \text{w.p. } z_2, \end{cases} \tag{14}$$

where

$$y_2 = \begin{cases} \Pi\Big(\frac{\gamma-\rho}{(1-p_{\text{err}})^{n-m} + p_{\text{err}}^{n-m}}\Big) & \beta \leqslant p_{\text{err}}^{n-m} \\ 0 & \beta > p_{\text{err}}^{n-m} \end{cases} \qquad z_2 = \Pi\Big(\frac{p_{\text{err}}^{n-m} - \beta}{(1-p_{\text{err}})^{n-m} + p_{\text{err}}^{n-m}}\Big)$$

$$\alpha \coloneqq (1 - p_{\text{attack}}) \sum_{k=0}^{\frac{n-1}{2}} \binom{n}{k} p_{\text{err}}^{n-k}(1-p_{\text{err}})^k \qquad \rho \coloneqq \sum_{k=m}^{\frac{n-1}{2}} \binom{n-m}{k-m} p_{\text{err}}^{n-k}(1-p_{\text{err}})^{k-m}$$

$$\gamma \coloneqq \sum_{k=0}^{\frac{n-1}{2}} \binom{n-m}{k} p_{\text{err}}^{n-m-k}(1-p_{\text{err}})^k \qquad \beta \coloneqq \frac{1-p_{\text{attack}}}{p_{\text{attack}}}\big((1-p_{\text{err}})^n - p_{\text{err}}^n\big)$$

and $\Pi : \mathbb{R} \to \mathbb{R}$ denotes the projection function into the interval $[0, 1]$:

$$\Pi(x) = \begin{cases} 0 & x < 0 \\ x & x \in [0,1] \\ 1 & x > 1. \end{cases}$$

□

1) Discussion: Conveniently, the optimal policy (13) for the detector is largely independent of the attack probability p_{attack}, whose value may be difficult to know precisely. In essence, the detector's policy only depends on p_{attack} because of the threshold condition that defines y_2:

$$\beta \coloneqq \frac{1-p_{\text{attack}}}{p_{\text{attack}}}\big((1-p_{\text{err}})^n - p_{\text{err}}^n\big) > p_{\text{err}}^{n-m}.$$

Moreover, for $m \geqslant 2$ and probabilities of attack satisfying (11), the previous inequality holds true for small values of p_{err} and the saddle-point for the detector only uses the majority rule.

While the detector's policy may depend little on p_{attack}, that is obviously not the case for the probability v^* of an estimation error corresponding to the saddle-point solution (12). For example, for very small probabilities of error, the saddle point is essentially given by

$$v^* \approx p_{\text{attack}} \binom{n-m}{\frac{n-1}{2}} p_{\text{err}}^{\frac{n+1}{2}-m},$$

which shows that the probability of an estimation error scales linearly with the attack probability. This formula also shows that the probability of an estimation error scales with the number of sensors as

$$p_{\text{err}}^{\frac{n+1-2m}{2}}. \tag{15}$$

In the absence of attacks (for which the majority rule would be optimal), we can conclude from Lemma 1 that the probability of an estimation error is given by

$$\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) = \sum_{k=\frac{n+1}{2}}^{n} \binom{n}{k} p_{\text{err}}^k (1-p_{\text{err}})^{n-k},$$

which, for a small probability p_{err} of sensor error, scales with the number of sensors as

$$p_{\text{err}}^{\frac{n+1}{2}}. \tag{16}$$

From the perspective of the scaling laws (15) and (16), it is as if each one of the m sensors compromised effectively decreases the total number of sensors by $2m$.

2) Proof of Theorem 1: The following proposition (proved in the appendix) is needed to prove Theorem 1.

Proposition 1: Given an integer $n > 2$, for every integers k and ℓ such that $1 \leqslant k \leqslant n-1$,

$$\binom{n}{k}\big(p_{\text{err}}^k(1-p_{\text{err}})^{n-k} - p_{\text{err}}^{n-k}(1-p_{\text{err}})^k\big) \;\leqslant\; n\big(p_{\text{err}}(1-p_{\text{err}})^{n-1} - p_{\text{err}}^{n-1}(1-p_{\text{err}})\big), \tag{17}$$

$\forall p_{\text{err}} \in (0, 2/(n+1)]$. □

Proof of Theorem 1. Using (6) and (7) in (5) one obtains

$$\begin{aligned}\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) = \frac{1}{2}\Bigg(& p_{\text{attack}} \sum_{k=m-\delta(0)}^{n-\delta(0)} \mu(k) \binom{n-m}{k-m+\delta(0)} p_{\text{err}}^{k-m+\delta(0)} (1-p_{\text{err}})^{n-k-\delta(0)} \\ &+ (1-p_{\text{attack}}) \sum_{k=0}^{n} \mu(k) \binom{n}{k} p_{\text{err}}^k (1-p_{\text{err}})^{n-k}\end{aligned}$$

$$+ p_{\text{attack}} \sum_{k=\delta(0)}^{n-m+\delta(0)} \mu(n-k) \binom{n-m}{k-\delta(0)} (1-p_{\text{err}})^{k-\delta(0)} p_{\text{err}}^{n-m-k+\delta(0)}$$
$$+ (1-p_{\text{attack}}) \sum_{k=0}^{n} \mu(n-k) \binom{n}{k} (1-p_{\text{err}})^{k} p_{\text{err}}^{n-k} \Big).$$

Making the change of variable $n-k \to \ell$ in the 3rd and 4th summations above, we obtain

$$\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) = p_{\text{attack}} \sum_{k=m-\delta(0)}^{n-\delta(0)} \mu(k) \binom{n-m}{k-m+\delta(0)} p_{\text{err}}^{k-m+\delta(0)} (1-p_{\text{err}})^{n-k-\delta(0)}$$
$$+ (1-p_{\text{attack}}) \sum_{k=0}^{n} \mu(k) \binom{n}{k} p_{\text{err}}^{k} (1-p_{\text{err}})^{n-k}.$$

Using the fact that n is odd and (6), we can break the summations as follows

$$\begin{aligned}
\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) =& p_{\text{attack}} \sum_{k=m-\delta(0)}^{\frac{n-1}{2}} \mu(k) \binom{n-m}{k-m+\delta(0)} p_{\text{err}}^{k-m+\delta(0)} (1-p_{\text{err}})^{n-k-\delta(0)} \\
&+ p_{\text{attack}} \sum_{k=\frac{n+1}{2}}^{n-\delta(0)} \big(1-\mu(n-k)\big) \binom{n-m}{k-m+\delta(0)} p_{\text{err}}^{k-m+\delta(0)} (1-p_{\text{err}})^{n-k-\delta(0)} \\
&+ (1-p_{\text{attack}}) \sum_{k=0}^{\frac{n-1}{2}} \mu(k) \binom{n}{k} p_{\text{err}}^{k} (1-p_{\text{err}})^{n-k} \\
&+ (1-p_{\text{attack}}) \sum_{k=\frac{n+1}{2}}^{n} \big(1-\mu(n-k)\big) \binom{n}{k} p_{\text{err}}^{k} (1-p_{\text{err}})^{n-k} \\
=& p_{\text{attack}} \sum_{k=m-\delta(0)}^{\frac{n-1}{2}} \mu(k) \binom{n-m}{k-m+\delta(0)} p_{\text{err}}^{k-m+\delta(0)} (1-p_{\text{err}})^{n-k-\delta(0)} \\
&+ p_{\text{attack}} \sum_{k=\delta(0)}^{\frac{n-1}{2}} \binom{n-m}{n-m-k+\delta(0)} p_{\text{err}}^{n-m-k+\delta(0)} (1-p_{\text{err}})^{k-\delta(0)} \\
&- p_{\text{attack}} \sum_{k=\delta(0)}^{\frac{n-1}{2}} \mu(k) \binom{n-m}{n-m-k+\delta(0)} p_{\text{err}}^{n-m-k+\delta(0)} (1-p_{\text{err}})^{k-\delta(0)} \\
&+ (1-p_{\text{attack}}) \sum_{k=0}^{\frac{n-1}{2}} \mu(k) \binom{n}{k} \Big(p_{\text{err}}^{k} (1-p_{\text{err}})^{n-k} - p_{\text{err}}^{n-k} (1-p_{\text{err}})^{k} \Big) + \alpha. \quad (18)
\end{aligned}$$

With this formula, we can proceed to exclude some of the estimation policies μ based on policy domination. To do this, we compute the derivative of the probability of an estimation error with respect to the values of $\mu(k)$. When this derivative is positive for every attack policy δ, we know

that we can restrict our attention to estimation policies for which $\mu(k) = 0$ since the policies with $\mu(k) = 1$ would be dominated (recall that the detector is the minimizer).

We consider separately four cases that differ by which summations in (18) include specific values of k:

1) For k such that $k \leqslant \min\{m - \delta(0), \delta(0)\}$ and $1 \leqslant k \leqslant \frac{n-1}{2}$, we have

$$\begin{aligned}\frac{d\,\mathrm{P}_{\mu,\delta}(\hat{X} \neq X)}{d\mu(k)} &= (1 - p_{\text{attack}})\binom{n}{k}\Big(p_{\text{err}}^{k}(1 - p_{\text{err}})^{n-k} - p_{\text{err}}^{n-k}(1 - p_{\text{err}})^{k}\Big) \\ &= (1 - p_{\text{attack}})\binom{n}{k}p_{\text{err}}^{k}(1 - p_{\text{err}})^{n-k}\Big(1 - \frac{p_{\text{err}}^{n-2k}}{(1 - p_{\text{err}})^{n-2k}}\Big) \geqslant 0,\end{aligned}$$

where the last inequality is a consequence of the fact that $p_{\text{err}} \leqslant 1/2$ and $n - 2k \geqslant 1$.

2) For k such that $m - \delta(0) \leqslant k < \delta(0)$ and $1 \leqslant k \leqslant \frac{n-1}{2}$, we have

$$\begin{aligned}\frac{d\,\mathrm{P}_{\mu,\delta}(\hat{X} \neq X)}{d\mu(k)} &= p_{\text{attack}}\binom{n-m}{k-m+\delta(0)}p_{\text{err}}^{k-m+\delta(0)}(1 - p_{\text{err}})^{n-k-\delta(0)} \\ &\quad + (1 - p_{\text{attack}})\binom{n}{k}\Big(p_{\text{err}}^{k}(1 - p_{\text{err}})^{n-k} - p_{\text{err}}^{n-k}(1 - p_{\text{err}})^{k}\Big) \geqslant 0,\end{aligned}$$

where the last inequality is again a consequence of the fact that $p_{\text{err}} \leqslant 1/2$ and $n-2k \geqslant 1$.

3) For k such that $\max\{m - \delta(0), \delta(0)\} \leqslant k$ and $1 \leqslant k \leqslant \frac{n-1}{2}$, we have

$$\begin{aligned}\frac{d\,\mathrm{P}_{\mu,\delta}(\hat{X} \neq X)}{d\mu(k)} &= p_{\text{attack}}\bigg(\binom{n-m}{k-m+\delta(0)}p_{\text{err}}^{k-m+\delta(0)}(1 - p_{\text{err}})^{n-k-\delta(0)} \\ &\qquad - \binom{n-m}{k-\delta(0)}p_{\text{err}}^{n-m-k+\delta(0)}(1 - p_{\text{err}})^{k-\delta(0)}\bigg) \\ &\quad + (1 - p_{\text{attack}})\binom{n}{k}\Big(p_{\text{err}}^{k}(1 - p_{\text{err}})^{n-k} - p_{\text{err}}^{n-k}(1 - p_{\text{err}})^{k}\Big) \\ &= p_{\text{attack}}\binom{n-m}{k-\delta(0)}p_{\text{err}}^{n-m-k+\delta(0)}(1 - p_{\text{err}})^{k-\delta(0)} \\ &\qquad \Big(\frac{(k-\delta(0))!(n-m-k+\delta(0))!}{(k-m+\delta(0))!(n-k-\delta(0))!}\Big(\frac{1 - p_{\text{err}}}{p_{\text{err}}}\Big)^{n-2k} - 1\Big) \\ &\quad + (1 - p_{\text{attack}})\binom{n}{k}\Big(p_{\text{err}}^{k}(1 - p_{\text{err}})^{n-k} - p_{\text{err}}^{n-k}(1 - p_{\text{err}})^{k}\Big) \\ &\geqslant p_{\text{attack}}\binom{n-m}{k-\delta(0)}p_{\text{err}}^{n-m-k+\delta(0)}(1 - p_{\text{err}})^{k-\delta(0)} \\ &\qquad \Big(\frac{\frac{n-2m+1}{2}!}{\frac{n-1}{2}!(n-m)!}\Big(\frac{1 - p_{\text{err}}}{p_{\text{err}}}\Big)^{n-2k} - 1\Big)\end{aligned}$$

$$
\begin{aligned}
&+ (1-p_{\text{attack}})\binom{n}{k}\Big(p_{\text{err}}^{k}(1-p_{\text{err}})^{n-k} - p_{\text{err}}^{n-k}(1-p_{\text{err}})^{k}\Big)\\
&\geqslant (1-p_{\text{attack}})n\Big(p_{\text{err}}(1-p_{\text{err}})^{n-1} - p_{\text{err}}^{n-1}(1-p_{\text{err}})\Big)\\
&= (1-p_{\text{attack}})np_{\text{err}}^{n-1}(1-p_{\text{err}})\Big(\Big(\frac{1-p_{\text{err}}}{p_{\text{err}}}\Big)^{n-2} - 1\Big) \geqslant 0,
\end{aligned}
$$

where the first inequality is a consequence of the facts that $k-\delta(0) \geqslant 0$, $k-m+\delta(0) \leqslant k \leqslant \frac{n-1}{2}$, $n-m-k+\delta(0) \geqslant n-m-\frac{n-1}{2} = \frac{n-2m+1}{2}$, $n-k-\delta(0) \leqslant n-m$; the second inequality a consequence of the fact that (10) is equivalent to

$$
\frac{\frac{n-2m+1}{2}!}{\frac{n-1}{2}!(n-m)!}\frac{1-p_{\text{err}}}{p_{\text{err}}} \geqslant 1
$$

which implies that

$$
\frac{\frac{n-2m+1}{2}!}{\frac{n-1}{2}!(n-m)!}\Big(\frac{1-p_{\text{err}}}{p_{\text{err}}}\Big)^{n-2k} \geqslant 1,\ \forall k \leqslant \frac{n-1}{2};
$$

the third inequality is a consequence of (17), which is valid in view of (9); and the last inequality is then a consequence of the fact that $p_{\text{err}} \leqslant 1/2$ and $n > 2$.

4) When $m \geqslant 2$, we also need to consider the case $\delta(0) \leqslant k < m-\delta(0)$ and $1 \leqslant k \leqslant \frac{n-1}{2}$, we have

$$
\begin{aligned}
\frac{d\,\mathrm{P}_{\mu,\delta}(\hat{X} \neq X)}{d\mu(k)} &= -\,p_{\text{attack}}\binom{n-m}{k-\delta(0)}p_{\text{err}}^{n-m-k+\delta(0)}(1-p_{\text{err}})^{k-\delta(0)}\\
&\quad + (1-p_{\text{attack}})\binom{n}{k}\Big(p_{\text{err}}^{k}(1-p_{\text{err}})^{n-k} - p_{\text{err}}^{n-k}(1-p_{\text{err}})^{k}\Big)\\
&\geqslant -\,p_{\text{attack}}\binom{n-m}{k-\delta(0)}p_{\text{err}}^{n-m-k+\delta(0)}(1-p_{\text{err}})^{k-\delta(0)}\\
&\quad + (1-p_{\text{attack}})n\Big(p_{\text{err}}(1-p_{\text{err}})^{n-1} - p_{\text{err}}^{n-1}(1-p_{\text{err}})\Big),
\end{aligned}
$$

where the inequality is a consequence of (17), which is valid in view of (9). Since we are dealing with a case for which $k \leqslant m-1$ we have that $k-\delta(0) \leqslant m-1$ and, because of (8), we have that

$$
k-\delta(0) \leqslant m-1 \leqslant (n-m)(1-p_{\text{err}}).
$$

Because of this, we can use the monotonicity of the binomial distribution up to its average, to conclude that

$$
\frac{d\,\mathrm{P}_{\mu,\delta}(\hat{X} \neq X)}{d\mu(k)} \geqslant -p_{\text{attack}}\binom{n-m}{m-1}p_{\text{err}}^{n-2m+1}(1-p_{\text{err}})^{m-1}
$$

$$+ (1 - p_{\text{attack}})n\Big(p_{\text{err}}(1 - p_{\text{err}})^{n-1} - p_{\text{err}}^{n-1}(1 - p_{\text{err}})\Big) \geqslant 0,$$

where the last inequality is then a consequence of (11).

The previous inequalities allow us to conclude that we only need to consider estimation policies with $\mu(k) = 0$ for all values of k that satisfy $1 \leqslant k \leqslant \frac{n-1}{2}$. We are thus left with only two estimation policies: the majority rule ($\mu(0) = 0$) and the no-consensus rule ($\mu(0) = 1$). For these pure policies, the probability of an estimation error (18) simplifies as follows: When $\delta(0) = 0$ (which corresponds to the deception rule) we have

$$\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) = \mu(0)p_{\text{attack}}(\beta - p_{\text{err}}^{n-m}) + p_{\text{attack}}\gamma + \alpha,$$

when $\delta(0) = m$ (which corresponds to the no-deception rule) we have

$$\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) = \mu(0)p_{\text{attack}}\big(\beta + (1 - p_{\text{err}})^{n-m}\big) + p_{\text{attack}}\rho + \alpha, \tag{19}$$

and when $0 < \delta(0) < m$ we have

$$\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) = \mu(0)p_{\text{attack}}\beta + \alpha. \tag{20}$$

Comparing (19) with (20), we conclude that the no-deception rule leads to a higher probability of an estimation error than any policy with $0 < \delta(0) < m$, and therefore the former dominates the latter. We are thus left, with the following 2×2 zero-sum game where the first row corresponds to the majority rule ($\mu(0) = 0$), the second row to the no-consensus rule ($\mu(0) = 1$), the first column to the deception rule ($\delta(0) = 0$), and the second column to the no-deception rule ($\delta(0) = m$):

$$\tilde{A} \coloneqq p_{\text{attack}} \begin{bmatrix} \gamma & \rho \\ \beta + \gamma - p_{\text{err}}^{n-m} & \beta + \rho + (1 - p_{\text{err}})^{n-m} \end{bmatrix} + \alpha \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

It is now straightforward to show that this matrix has a mixed saddle-point

$$y^* \coloneqq \begin{bmatrix} 1 - y_2 \\ y_2 \end{bmatrix}, \qquad\qquad z^* \coloneqq \begin{bmatrix} 1 - z_2 \\ z_2 \end{bmatrix},$$

with value v^* (cf., e.g., [3]). ■

IV. Approximate saddle-point

In view of the condition (9), the estimation policy provided by Theorem 1 is only optimal for a number of sensors n roughly below $2/p_{\text{err}}$. We shall see in this section that when the number of sensors is large and the probability of sensor error is not very small, a threshold estimation policy like the majority rule is almost optimal. However, the proof of this result requires a completely different approach, which is described next.

The random variable

$$\bar{Z} \coloneqq \sum_{i=1}^{n} Z_i,$$

observed by the detector may have different distributions depending on the value of X and whether or not there is an attack. Specifically,

$$\bar{Z} = \begin{cases} R & \text{w.p. } 1 - p_{\text{attack}}, \\ S + W & \text{w.p. } p_{\text{attack}}, \end{cases} \tag{21}$$

where the random variable R equals the sum of all the Y_i and therefore its distribution is

$$R \sim \begin{cases} \text{Binom}(n, p_{\text{err}}) & X = 0, \\ \text{Binom}(n, 1 - p_{\text{err}}) & X = 1; \end{cases} \tag{22}$$

the random variable S equals the sum of the $n - m$ sensors that have not been compromised by the attacker and therefore has distribution

$$S \sim \begin{cases} \text{Binom}(n - m, p_{\text{err}}) & X = 0, \\ \text{Binom}(n - m, 1 - p_{\text{err}}) & X = 1; \end{cases} \tag{23}$$

and the random variable W equals the sum of the readings of the m sensors compromised by the attacker. The distribution of W is selected by the attacker and may depend on the value of X, with the constraint that its support must lie in the set $\{0, 1, \ldots, m\}$. This perspective motivates the general problem formulated and solved in the next sections.

A. General case

Suppose that one wants to estimate the random variables X with Bernoulli distribution (1) based on a measurement $\bar{Z}$ of the form (21) where the conditional distributions of R and S given X are known and the conditional distribution of W given X is selected by an adversary,

but its support is limited to a given subset $I \subset \mathbb{R}$. As before, we formulate this as a zero-sum game where the estimator wants to minimize the probability of an estimation error, whereas the attacker want to maximize this probability.

We allow the *estimation policy* to be stochastic and represent it by a function $f : \mathbb{R} \to [0, 1]$, with the understanding that, when the estimator observes a value $\bar{z} \in \mathbb{R}$ for (21), she selects

$$\hat{X} = \begin{cases} 1 & \text{w.p. } f(\bar{z}) \\ 0 & \text{w.p. } 1 - f(\bar{z}). \end{cases}$$

Denoting by ρ_x, σ_x, and ω_x, respectively, the conditional distributions of R, S, and W given that $X = x$, we can use the law of total probability (much like in the proof of Lemma 1), to express the probability of an estimation error as follows:

$$\begin{aligned}
&J(f, \omega_0, \omega_1) \coloneqq \mathrm{P}(\hat{\mu} \neq \mu) \\
&\quad = p_{\text{attack}}(1-p)\,\mathrm{P}(\hat{X} = 1 \mid X = 0, \mathcal{E}_{\text{attack}}) + p_{\text{attack}}\,p\,\mathrm{P}(\hat{X} = 0 \mid X = 1, \mathcal{E}_{\text{attack}}) \\
&\qquad (1 - p_{\text{attack}})(1-p)\,\mathrm{P}(\hat{X} = 1 \mid X = 0, \neg\mathcal{E}_{\text{attack}}) + (1 - p_{\text{attack}})\,p\,\mathrm{P}(\hat{X} = 0 \mid X = 1, \neg\mathcal{E}_{\text{attack}}) \\
&\quad = p_{\text{attack}}(1-p) \int_I \int_{\mathbb{R}} f(\bar{y} + \bar{w})\sigma_0(d\bar{y})\omega_0(d\bar{w}) + p_{\text{attack}}\,p \int_I \int_{\mathbb{R}} \big(1 - f(\bar{y} + \bar{w})\big)\sigma_1(d\bar{y})\omega_1(d\bar{w}) \\
&\qquad + (1 - p_{\text{attack}})(1-p) \int_{\mathbb{R}} f(\bar{y})\rho_0(d\bar{y}) + (1 - p_{\text{attack}})\,p \int_{\mathbb{R}} \big(1 - f(\bar{y})\big)\rho_1(d\bar{y}).
\end{aligned}$$

where $\mathcal{E}_{\text{attack}}$ denotes the event that the attacker manipulated measurements. Grouping all the terms that do not depend on f, the above expression can be simplified to

$$\begin{aligned}
J(f, \omega_0, \omega_1) = p &+ p_{\text{attack}}(1-p) \int_I \int_{\mathbb{R}} f(\bar{y} + \bar{w})\sigma_0(d\bar{y})\omega_0(d\bar{w}) \\
&- p_{\text{attack}}\,p \int_I \int_{\mathbb{R}} f(\bar{y} + \bar{w})\sigma_1(d\bar{y})\omega_1(d\bar{w}) \\
&+ (1 - p_{\text{attack}})(1-p) \int_{\mathbb{R}} f(\bar{y})\rho_0(d\bar{y}) - (1 - p_{\text{attack}})p \int_{\mathbb{R}} f(\bar{y})\rho_1(d\bar{y}). \qquad (24)
\end{aligned}$$

For the above formula to be well defined, we assume that the attacker is only allowed to select distributions (ω_0, ω_1) in a set $\mathcal{A}$ containing all pairs of distributions (ω_0, ω_1) for which the integrals in (24) exist for every Lebesgue measurable function f.

The problem just defined matches exactly the one considered in Section II when $\rho_0, \rho_1, \sigma_0, \sigma_1$ are the binomial distributions defined by (22)–(23) and $I \coloneqq \{0, 1, 2, \ldots, m\}$. However, we start by computing saddle point policies for the simpler case where $\rho_0, \rho_1, \sigma_0, \sigma_1$ are continuous distributions and I is an interval.

B. Continuous-distributions case

When the set I is an interval, the following result provides a sufficient condition for the problem defined above to have a saddle point solution for which ω_0, ω_1 are Dirac distributions at the two extrema of I. While this condition may seem very restrictive, we shall see shortly that it holds when $\rho_0, \rho_1, \sigma_0, \sigma_1$ are Gaussian distributions.

Theorem 2: Suppose that $I = [a, b]$, that $\rho_0, \rho_1, \sigma_0, \sigma_1$ are continuous distributions (with probability density functions), and define the set

$$\mathcal{Z} \coloneqq \{\bar{z} \in \mathbb{R} : g(\bar{z}) \leqslant 0\}, \tag{25}$$

where[1]

$$g(\bar{z}) \coloneqq p_{\text{attack}}\Big((1-p)\sigma_0(\bar{z}-b) - p\sigma_1(\bar{z}-a)\Big) + (1-p_{\text{attack}})\Big((1-p)\rho_0(\bar{z}) - p\rho_1(\bar{z})\Big). \tag{26}$$

Assuming that

$$\int_{\mathcal{Z}} \sigma_0(\bar{z}-\bar{w})d\bar{z} \leqslant \int_{\mathcal{Z}} \sigma_0(\bar{z}-b)d\bar{z}, \tag{27}$$

$$\int_{\mathcal{Z}} \sigma_1(\bar{z}-\bar{w})d\bar{z} \geqslant \int_{\mathcal{Z}} \sigma_1(\bar{z}-a)d\bar{z}, \tag{28}$$

$\forall \bar{w} \in I$, then the function

$$f^*(\bar{z}) \coloneqq \begin{cases} 0 & \bar{z} \in \mathbb{R}\backslash\mathcal{Z} \\ 1 & \bar{z} \in \mathcal{Z} \end{cases}$$

for the defender and the functions

$$\omega_0^*(\bar{w}) = \boldsymbol{\delta}(\bar{w}-b), \qquad \omega_1^*(\bar{w}) = \boldsymbol{\delta}(\bar{w}-a), \tag{29}$$

for the attacker form a saddle-point for the game with value given by

$$J(f^*, \omega_0^*, \omega_1^*) = p + \int_{\mathcal{Z}} g(\bar{z})d\bar{z}. \qquad \square$$

Proof of Theorem 2. When the distributions $\rho_0, \rho_1, \sigma_0, \sigma_1$ are continuous (and have pdfs), the probability of an estimation error (24) can be re-written as

$$J(f,\omega_0, \omega_1) = p + p_{\text{attack}}(1-p)\int_{\mathbb{R}}\int_I f(\bar{y}+\bar{w})\sigma_0(\bar{y})\omega_0(d\bar{w})d\bar{y}$$

[1]With some abuse of notation, we use here the same symbols $\rho_0, \rho_1, \sigma_0, \sigma_1$ for the continuous distributions and for their probability density functions (pdfs).

$$
\begin{aligned}
&- p_{\text{attack}}\, p \int_{\mathbb{R}} \int_{I} f(\bar{y}+\bar{w})\sigma_1(\bar{y})\omega_1(d\bar{w})d\bar{y} \\
&+ (1-p_{\text{attack}})(1-p)\int_{\mathbb{R}} f(\bar{y})\rho_0(\bar{y})d\bar{y} - (1-p_{\text{attack}})\, p \int_{\mathbb{R}} f(\bar{y})\rho_1(\bar{y})d\bar{y} \\
&= p + p_{\text{attack}}(1-p)\int_{\mathbb{R}} \int_{I} f(\bar{z})\sigma_0(\bar{z}-\bar{w})\omega_0(d\bar{w})d\bar{z} - p_{\text{attack}}\, p \int_{\mathbb{R}} \int_{I} f(\bar{z})\sigma_1(\bar{z}-\bar{w})\omega_1(d\bar{w})d\bar{z} \\
&+ (1-p_{\text{attack}})(1-p)\int_{\mathbb{R}} f(\bar{y})\rho_0(\bar{y})d\bar{y} - (1-p_{\text{attack}})\, p \int_{\mathbb{R}} f(\bar{y})\rho_1(\bar{y})d\bar{y}. \qquad (30)
\end{aligned}
$$

We prove this theorem by showing that the given policies satisfy the following saddle-point inequalities:

$$J(f^*, \omega_0, \omega_1) \leqslant J(f^*, \omega_0^*, \omega_1^*) \leqslant J(f, \omega_0^*, \omega_1^*)$$

for every Lebesgue measurable function f and for every pair of functions $(\omega_0, \omega_1) \in \mathcal{A}$. Suppose first that the attacker selects the distributions (29). In this case, for an arbitrary Lebesgue measurable function f, (30) leads to

$$
\begin{aligned}
J(f, \omega_0^*, \omega_1^*) &= p + p_{\text{attack}}(1-p)\int_{\mathbb{R}} f(\bar{z})\sigma_0(\bar{z}-b)d\bar{z} - p_{\text{attack}}\, p \int_{\mathbb{R}} f(\bar{z})\sigma_1(\bar{z}-a)d\bar{z} \\
&\quad + (1-p_{\text{attack}})(1-p)\int_{\mathbb{R}} f(\bar{y})\rho_0(\bar{y})d\bar{y} - (1-p_{\text{attack}})\, p \int_{\mathbb{R}} f(\bar{y})\rho_1(\bar{y})d\bar{y} \\
&= p + \int_{\mathbb{R}} f(\bar{z})g(\bar{z})d\bar{z} \\
&\geqslant \inf_f \Big(p + \int_{\mathbb{R}} f(\bar{z})g(\bar{z})d\bar{z} \Big) = p + \int_{\mathcal{Z}} g(\bar{z})d\bar{z} = J(f^*, \omega_0^*, \omega_1^*), \qquad (31)
\end{aligned}
$$

where the infimum is taken over Lebesgue measurable functions taking values in $[0, 1]$ and is achieved for the function f^* that takes the value 1 when $g(\bar{z}) < 0$ and 0 otherwise.

Suppose now that the detector uses the function f^*. For arbitrary distributions $(\omega_0, \omega_1) \in \mathcal{A}$, conclude from (30) that

$$
\begin{aligned}
J(f^*,&\omega_0, \omega_1) = p + p_{\text{attack}}(1-p)\int_{I}\int_{\mathcal{Z}} \sigma_0(\bar{z}-\bar{w})d\bar{z}\omega_0(d\bar{w}) - p_{\text{attack}}\, p \int_{I}\int_{\mathcal{Z}} \sigma_1(\bar{z}-\bar{w})d\bar{z}\omega_1(d\bar{w}) \\
&\quad + (1-p_{\text{attack}})(1-p)\int_{\mathcal{Z}} \rho_0(\bar{y})d\bar{y} - (1-p_{\text{attack}})\, p \int_{\mathcal{Z}} \rho_1(\bar{y})d\bar{y} \\
&\leqslant p + p_{\text{attack}}(1-p)\int_{I}\int_{\mathcal{Z}} \sigma_0(\bar{z}-b)d\bar{z}\omega_0(d\bar{w}) - p_{\text{attack}}\, p \int_{I}\int_{\mathcal{Z}} \sigma_1(\bar{z}-a)d\bar{z}\omega_1(d\bar{w}) \\
&\quad + (1-p_{\text{attack}})(1-p)\int_{\mathcal{Z}} \rho_0(\bar{y})d\bar{y} - (1-p_{\text{attack}})\, p \int_{\mathcal{Z}} \rho_1(\bar{y})d\bar{y} \\
&= p + p_{\text{attack}}(1-p)\int_{\mathcal{Z}} \sigma_0(\bar{z}-b)d\bar{z} - p_{\text{attack}}\, p \int_{\mathcal{Z}} \sigma_1(\bar{z}-a)d\bar{z}
\end{aligned}
$$

$$+ (1 - p_{\text{attack}})(1 - p) \int_{\mathcal{Z}} \rho_0(\bar{y}) d\bar{y} - (1 - p_{\text{attack}})\, p \int_{\mathcal{Z}} \rho_1(\bar{y}) d\bar{y}$$
$$= p + \int_{\mathcal{Z}} g(\bar{z}) d\bar{z} = J(f^*, \omega_0^*, \omega_1^*),$$

where the inequality is a consequence of (27)–(28) and the final equalities result from the definition of g and (31), respectively. ■

C. Gaussian case

When $\rho_0, \rho_1, \sigma_0, \sigma_1$ are Gaussian distributions one can obtain explicit formulas for the saddle-point policies in Theorem 2.

Corollary 1: Suppose that $I = [a, b]$, that each ρ_x, $x \in \{0, 1\}$ is a normal distribution with mean $\bar{\rho}_x$ and variance σ_ρ^2, and that each σ_x, $x \in \{0, 1\}$ is a normal distribution with mean $\bar{\sigma}_x$ and variance σ_σ^2. Assuming that

$$\max\left\{\frac{b + \bar{\sigma}_0}{\sigma_\sigma^2}, \frac{\bar{\rho}_0}{\sigma_\rho^2}\right\} < \min\left\{\frac{a + \bar{\sigma}_1}{\sigma_\sigma^2}, \frac{\bar{\rho}_1}{\sigma_\rho^2}\right\}, \tag{32}$$

and that σ_σ^2 is sufficiently close to σ_ρ^2, then the function

$$g(\bar{z}) \coloneqq \frac{p_{\text{attack}}}{\sigma_\sigma \sqrt{2\pi}}\left((1-p)e^{-\frac{(\bar{z}-b-\bar{\sigma}_0)^2}{2\sigma_\sigma^2}} - pe^{-\frac{(\bar{z}-a-\bar{\sigma}_1)^2}{2\sigma_\sigma^2}}\right) + \frac{1 - p_{\text{attack}}}{\sigma_\rho \sqrt{2\pi}}\left((1-p)e^{-\frac{(\bar{z}-\bar{\rho}_0)^2}{2\sigma_\rho^2}} - pe^{-\frac{(\bar{z}-\bar{\rho}_1)^2}{2\sigma_\rho^2}}\right), \tag{33}$$

has a unique zero $\bar{z} = z^*$, the set $\mathcal{Z}$ in (25) is of the form

$$\mathcal{Z} \coloneqq \{\bar{z} \in \mathbb{R} : \bar{z} \geqslant z^*\}, \tag{34}$$

the equations (27)–(28) hold, and therefore the function

$$f^*(\bar{z}) \coloneqq \begin{cases} 0 & \bar{z} < \bar{z}^* \\ 1 & \bar{z} \geqslant \bar{z}^* \end{cases} \tag{35}$$

for the defender and the functions

$$\omega_0^*(\bar{w}) = \boldsymbol{\delta}(\bar{w} - b), \qquad \omega_1^*(\bar{w}) = \boldsymbol{\delta}(\bar{w} - a),$$

for the attacker form a saddle-point with value given by

$$J(f^*, \omega_0^*, \omega_1^*) = p + \int_{z^*}^{\infty} g(\bar{z}) d\bar{z}.$$ □

Remark 1: The corollary's assumption that "σ_σ^2 is sufficiently close to σ_ρ^2," is only needed to make sure that the function (33) has a unique zero. One could construct an explicit bound on how close the two variances need to be for this to happen along the lines of the proof of Proposition 2 in the appendix, but we do not include such bound here because it is conservative and it is straightforward to verify numerically whether or not the function (33) has a unique zero. □

Proof of Corollary 1. For the given distributions the function (26) is given by (33), which is of the form (55) in the Proposition 2 in the appendix, for

$$a_2 = \frac{1}{2\sigma_\sigma^2}, \qquad b_1 = \frac{b+\bar{\sigma}_0}{\sigma_\sigma^2}, \qquad c_1 = \frac{a+\bar{\sigma}_1}{\sigma_\sigma^2},$$
$$\bar{a}_2 = \frac{1}{2\sigma_\rho^2}, \qquad \bar{b}_1 = \frac{\bar{\rho}_0}{\sigma_\rho^2}, \qquad \bar{c}_1 = \frac{\bar{\rho}_1}{\sigma_\rho^2}.$$

Since (32) guarantees that (56) holds, we conclude from Proposition 2 that for a_2 sufficiently close to $\bar{a}_2$, (59) has a unique zero $\bar{z} = z^*$ and that

$$g(\bar{z}) \leqslant 0 \quad \Leftrightarrow \quad \bar{z} \geqslant z^*,$$

leading to (34).

To verify the conditions (27)–(28), we write them for the given pdfs and the set $\mathcal{Z}$, leading to

$$\int_{z^*}^{\infty} e^{-\frac{(\bar{z}-\bar{w}-\bar{\sigma}_0)^2}{2\sigma_\sigma^2}} d\bar{z} \leqslant \int_{z^*}^{\infty} e^{-\frac{(\bar{z}-b-\bar{\sigma}_0)^2}{2\sigma_\sigma^2}} d\bar{z}, \qquad \int_{z^*}^{\infty} e^{-\frac{(\bar{z}-\bar{w}-\bar{\sigma}_1)^2}{2\sigma_\sigma^2}} d\bar{z} \geqslant \int_{z^*}^{\infty} e^{-\frac{(\bar{z}-a-\bar{\sigma}_1)^2}{2\sigma_\sigma^2}} d\bar{z},$$

(modulo a multiplication by $\sigma_\sigma\sqrt{2\pi} > 0$). Making appropriate changes of integration variables, we obtain the equivalent expressions

$$\int_{z^*-\bar{w}}^{\infty} e^{-\frac{(\bar{s}-\bar{\sigma}_0)^2}{2\sigma_\sigma^2}} d\bar{s} \leqslant \int_{z^*-b}^{\infty} e^{-\frac{(\bar{s}-\bar{\sigma}_0)^2}{2\sigma_\sigma^2}} d\bar{s}, \qquad \int_{z^*-\bar{w}}^{\infty} e^{-\frac{(\bar{s}-\bar{\sigma}_1)^2}{2\sigma_\sigma^2}} d\bar{s} \geqslant \int_{z^*-a}^{\infty} e^{-\frac{(\bar{s}-\bar{\sigma}_1)^2}{2\sigma_\sigma^2}} d\bar{s},$$

which indeed hold for every $\bar{w} \in [a, b]$. ■

D. Binomial case

When $\rho_0, \rho_1, \sigma_0, \sigma_1$ are the binomial distributions in (22)–(23), Theorem 2 no longer applies. However, since binomial distributions can be well approximated by Gaussian distributions, we shall see that it is possible to use Corollary 1 to compute an ϵ-saddle point for a small value

of ϵ. We recall that a pair $(u^*, d^*) \in \mathcal{U} \times \mathcal{D}$ is an ϵ-saddle-point with respect to a criterion $J : \mathcal{U} \times \mathcal{D} \to \mathbb{R}$ when

$$J(u^*, d^*) - \epsilon \leqslant J(u^*, d^*) \leqslant J(u, d^*) + \epsilon, \ \forall u \in \mathcal{U}, d \in \mathcal{D}.$$

For $\epsilon = 0$, an ϵ-saddle-point is just a regular saddle-point.

Theorem 3: Suppose that $\rho_0, \rho_1, \sigma_0, \sigma_1$ are the binomial distributions in (22)–(23), that $I = \{0, 1, \ldots, m\}$ with

$$m < n(1 - 2p_{\text{err}}), \quad p_{\text{err}} \in (0, 1/2), \tag{36}$$

and that the function

$$\begin{aligned} g(\bar{z}) \coloneqq & \frac{p_{\text{attack}}}{np_{\text{err}}(1-p_{\text{err}})\sqrt{2\pi}}\Big((1-p)e^{-\frac{(\bar{z}-m-np_{\text{err}})^2}{2np_{\text{err}}(1-p_{\text{err}})}} - pe^{-\frac{(\bar{z}-n(1-p_{\text{err}}))^2}{2np_{\text{err}}(1-p_{\text{err}})}}\Big) \\ & + \frac{1-p_{\text{attack}}}{(n-m)p_{\text{err}}(1-p_{\text{err}})\sqrt{2\pi}}\Big((1-p)e^{-\frac{(\bar{z}-(n-m)p_{\text{err}})^2}{2(n-m)p_{\text{err}}(1-p_{\text{err}})}} - pe^{-\frac{(\bar{z}-(n-m)(1-p_{\text{err}}))^2}{2(n-m)p_{\text{err}}(1-p_{\text{err}})}}\Big), \end{aligned} \tag{37}$$

has a unique zero $\bar{z} = z^*$. Then the function

$$f^*(\bar{z}) \coloneqq \begin{cases} 0 & \bar{z} < \bar{z}^* \\ 1 & \bar{z} \geqslant \bar{z}^* \end{cases} \tag{38}$$

for the defender and the functions

$$\omega_0^*(\bar{w}) = \boldsymbol{\delta}(\bar{w} - m), \qquad \omega_1^*(\bar{w}) = \boldsymbol{\delta}(\bar{w}), \tag{39}$$

for the attacker form an 2ϵ-saddle-point with value $J(f^*, \omega_0^*, \omega_1^*)$ satisfying

$$\Big|J(f^*, \omega_0^*, \omega_1^*) - p - \int_{z^*}^{\infty} g(\bar{z})d\bar{z}\Big| \leqslant \epsilon,$$

for

$$\begin{aligned} \epsilon \coloneqq & (1-p)\big(p_{\text{attack}}\mathcal{E}(n, p_{\text{err}}) + (1-p_{\text{attack}})\mathcal{E}(n-m, p_{\text{err}})\big) \\ & + p\big(p_{\text{attack}}\mathcal{E}(n, 1-p_{\text{err}}) + (1-p_{\text{attack}})\mathcal{E}(n-m, 1-p_{\text{err}})\big), \end{aligned} \tag{40}$$

with

$$\begin{aligned} \mathcal{E}(\ell, q) \coloneqq \max\Bigg\{ & 1 + \frac{1}{2}\operatorname{erf}\Big(\frac{\eta - \ell q}{\sqrt{2\ell q(1-q)}}\Big) - \frac{1}{2}\operatorname{erf}\Big(\frac{\ell + 1 + \eta - \ell q}{\sqrt{2\ell q(1-q)}}\Big) \\ & + \sum_{k=0}^{\ell} H\Bigg(\frac{1}{2}\operatorname{erf}\Big(\frac{k + \eta + 1 - \ell q}{\sqrt{2\ell q(1-q)}}\Big) - \frac{1}{2}\operatorname{erf}\Big(\frac{k + \eta - \ell q}{\sqrt{2\ell q(1-q)}}\Big) - \binom{\ell}{k} q^k (1-q)^{\ell-k}\Bigg), \end{aligned}$$

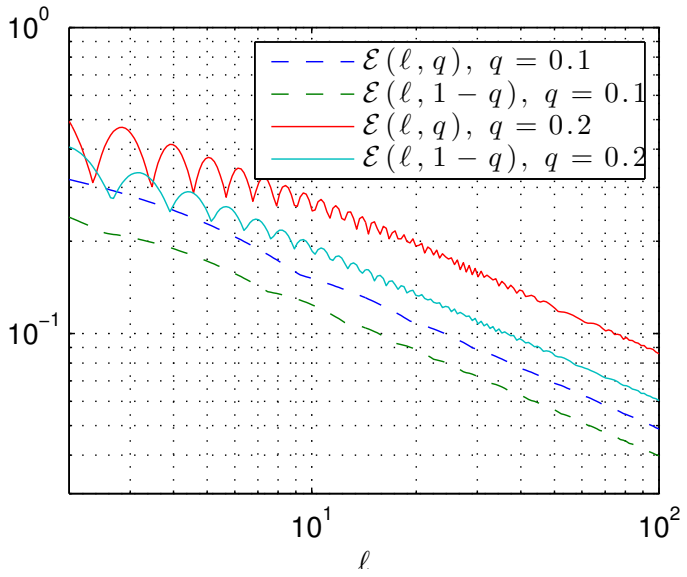

Fig. 2. Function $\mathcal{E}(\ell, q)$ defined in (41) for $\eta = 0$ (other values of η lead to smaller values).

$$- \sum_{k=0}^{\ell} H\left(\binom{\ell}{k} q^k (1-q)^{\ell-k} - \frac{1}{2} \operatorname{erf}\left(\frac{k+\eta+1-\ell q}{\sqrt{2\ell q(1-q)}} \right) + \frac{1}{2} \operatorname{erf}\left(\frac{k+\eta-\ell q}{\sqrt{2\ell q(1-q)}} \right) \right) \Bigg\}, \quad (41)$$

where $\operatorname{erf}(s) \coloneqq \frac{1}{\sqrt{\pi}} \int_{-s}^{s} e^{-t^2} dt$, η is equal to the distance between z^* and its closest integer, and H denotes the Heaviside step function $H(s) = 1$, $\forall s \geqslant 0$ and $H(s) = 0$, $\forall s < 0$. □

The function $\mathcal{E}(\ell, q)$ in (41) essentially provides an error between the probability that a Gaussian random variable with mean ℓq and variance $\ell q(1-q)$ falls between $k+\eta$ and $k+\eta+1$ and the probability that a binomial random variable with parameters ℓ and q take the value k. By the Moivre-Laplace theorem such errors decrease to zero as fast as $1/\sqrt{\ell}$ (see Figure 2) and therefore ϵ in (40) converges to zero as fast as $\sqrt{n-m}$.

The proof of Theorem 3 requires two results stated below. The first (Lemma 3) shows that if we replace in the criterion J defined by (24) the distributions $\rho_0, \rho_1, \sigma_0, \sigma_1$ by "similar" distributions $\tilde{\rho}_0, \tilde{\rho}_1, \tilde{\sigma}_0, \tilde{\sigma}_1$, then the resulting new criterion $\tilde{J}$ is "close" to J. The second result (Lemma 2) then shows that a saddle-point for a criterion J is an ϵ-saddle-point for another criterion $\tilde{J}$ that is "close" to J.

Lemma 2: Consider two zero-sum game criteria $J : \mathcal{U} \times \mathcal{D} \to \mathbb{R}$ and $\tilde{J} : \tilde{\mathcal{U}} \times \tilde{\mathcal{D}} \to \mathbb{R}$ with $\tilde{\mathcal{U}} \subset \mathcal{U}$ and $\tilde{\mathcal{D}} \subset \mathcal{D}$, such that

$$|J(u,d) - \tilde{J}(u,d)| \leqslant \epsilon, \qquad \forall u \in \tilde{\mathcal{U}}, d \in \tilde{\mathcal{D}}, \quad (42)$$

for a given $\epsilon \geqslant 0$. If a pair $(u^*, d^*) \in \tilde{\mathcal{U}} \times \tilde{\mathcal{D}} \subset \mathcal{U} \times \mathcal{D}$ is a saddle-point with respect to J, then the same pair (u^*, d^*) is a 2ϵ-saddle-point with respect to a criterion $\tilde{J}$. □

It should be emphasized that the error-bound in (42) only needs to hold on the (smaller) domain of $\tilde{J}$. The proof of this result is a straightforward consequence of the definitions of saddle-point and ϵ-saddle-point equilibria and can be found in the appendix.

Lemma 3: Consider the criteria J defined by (24) using four distributions $\rho_0, \rho_1, \sigma_0, \sigma_1$ and another criteria $\tilde{J}$ defined by

$$\begin{aligned}\tilde{J}(f, \omega_0, \omega_1) = p &+ p_{\text{attack}}(1-p) \int_I \int_{\mathbb{R}} f(\bar{y} + \bar{w}) \tilde{\sigma}_0(d\bar{y}) \omega_0(d\bar{w}) \\ &- p_{\text{attack}}\, p \int_I \int_{\mathbb{R}} f(\bar{y} + \bar{w}) \tilde{\sigma}_1(d\bar{y}) \omega_1(d\bar{w}) \\ &+ (1 - p_{\text{attack}})(1-p) \int_{\mathbb{R}} f(\bar{y}) \tilde{\rho}_0(d\bar{y}) - (1 - p_{\text{attack}}) p \int_{\mathbb{R}} f(\bar{y}) \tilde{\rho}_1(d\bar{y}), \quad (43)\end{aligned}$$

using four alternative distributions $\tilde{\rho}_0, \tilde{\rho}_1, \tilde{\sigma}_0, \tilde{\sigma}_1$. For every discrete distributions ω_0, ω_1 of the form

$$\omega_0(\bar{w}) = \sum_{i=-\infty}^{\infty} p_i \boldsymbol{\delta}(\bar{w} - i), \qquad \sum_{i=-\infty}^{\infty} p_i = 1 \quad (44)$$

$$\omega_1(\bar{w}) = \sum_{i=-\infty}^{\infty} q_i \boldsymbol{\delta}(\bar{w} - i), \qquad \sum_{i=-\infty}^{\infty} q_i = 1, \quad (45)$$

we have that

$$|J(f, \omega_0, \omega_1) - \tilde{J}(f, \omega_0, \omega_1)| \leqslant \epsilon,$$

where

$$\epsilon \coloneqq (1-p)\big(c_0 p_{\text{attack}} + d_0(1 - p_{\text{attack}})\big) + p\big(c_1 p_{\text{attack}} + d_1(1 - p_{\text{attack}})\big) \quad (46)$$

$$c_0 \coloneqq \sup_{i \in \mathbb{Z}} \Big| \int_{\mathbb{R}} f(\bar{y} + i)\big(\sigma_0(d\bar{y}) - \tilde{\sigma}_0(d\bar{y})\big)\Big|, \quad (47)$$

$$c_1 \coloneqq \sup_{i \in \mathbb{Z}} \Big| \int_{\mathbb{R}} f(\bar{y} + i)\big(\sigma_1(d\bar{y}) - \tilde{\sigma}_1(d\bar{y})\big)\Big|, \quad (48)$$

$$d_0 \coloneqq \Big| \int_{\mathbb{R}} f(\bar{y})\big(\rho_0(d\bar{y}) - \tilde{\rho}_0(d\bar{y})\big)\Big|, \quad (49)$$

$$d_1 \coloneqq \Big| \int_{\mathbb{R}} f(\bar{y})\big(\rho_1(d\bar{y}) - \tilde{\rho}_1(d\bar{y})\big)\Big|. \quad (50)$$

□

Proof of Lemma 3. The error $J - \tilde{J}$ can be written as

$$\begin{aligned}
J(f,\omega_0,\omega_1) - \tilde{J}(f,\omega_0,\omega_1) &= p_{\text{attack}}(1-p)\int_I\int_{\mathbb{R}} f(\bar{y}+\bar{w})\big(\sigma_0(d\bar{y}) - \tilde{\sigma}_0(d\bar{y})\big)\omega_0(d\bar{w}) \\
&\quad - p_{\text{attack}}\,p\int_I\int_{\mathbb{R}} f(\bar{y}+\bar{w})\big(\sigma_1(d\bar{y}) - \tilde{\sigma}_1(d\bar{y})\big)\omega_1(d\bar{w}) \\
&\quad + (1-p_{\text{attack}})(1-p)\int_{\mathbb{R}} f(\bar{y})\big(\rho_0(d\bar{y}) - \tilde{\rho}_0(d\bar{y})\big) - (1-p_{\text{attack}})p\int_{\mathbb{R}} f(\bar{y})\big(\rho_1(d\bar{y}) - \tilde{\rho}_1(d\bar{y})\big),
\end{aligned}$$

and therefore, for distributions of the form (44)–(45), we conclude that

$$\begin{aligned}
J(f,\omega_0,\omega_1) - \tilde{J}(f,\omega_0,\omega_1) &= p_{\text{attack}}(1-p)\sum_{i=-\infty}^{\infty} p_i\int_{\mathbb{R}} f(\bar{y}+i)\big(\sigma_0(d\bar{y}) - \tilde{\sigma}_0(d\bar{y})\big) \\
&\quad - p_{\text{attack}}\,p\sum_{i=-\infty}^{\infty} q_i\int_{\mathbb{R}} f(\bar{y}+i)\big(\sigma_1(d\bar{y}) - \tilde{\sigma}_1(d\bar{y})\big) \\
&\quad + (1-p_{\text{attack}})(1-p)\int_{\mathbb{R}} f(\bar{y})\big(\rho_0(d\bar{y}) - \tilde{\rho}_0(d\bar{y})\big) - (1-p_{\text{attack}})p\int_{\mathbb{R}} f(\bar{y})\big(\rho_1(d\bar{y}) - \tilde{\rho}_1(d\bar{y})\big).
\end{aligned}$$

Using (47)–(50) that the fact that $\sum_i p_i = \sum_i q_i = 1$, we then obtain

$$\begin{aligned}
|J(f,\omega_0,\omega_1) - \tilde{J}(f,\omega_0,\omega_1)| &\leqslant c_0 p_{\text{attack}}(1-p) + c_1 p_{\text{attack}}\,p + d_0(1-p_{\text{attack}})(1-p) + d_1(1-p_{\text{attack}})p \\
&= (1-p)\big(c_0 p_{\text{attack}} + d_0(1-p_{\text{attack}})\big) + p\big(c_1 p_{\text{attack}} + d_1(1-p_{\text{attack}})\big). \qquad \blacksquare
\end{aligned}$$

Proof of Theorem 3. To prove this result we consider two games: one *continuous game* defined by (24) with four Gaussian distributions

$$\begin{aligned}
\sigma_0 &\sim \mathcal{N}\big(\bar{\sigma}_0 \coloneqq np_{\text{err}}, {\sigma_\sigma}^2 \coloneqq np_{\text{err}}(1-p_{\text{err}})\big) \\
\sigma_1 &\sim \mathcal{N}\big(\bar{\sigma}_1 \coloneqq n(1-p_{\text{err}}), {\sigma_\sigma}^2\big) \\
\rho_0 &\sim \mathcal{N}\big(\bar{\rho}_0 \coloneqq (n-m)p_{\text{err}}, {\sigma_\rho}^2 \coloneqq (n-m)p_{\text{err}}(1-p_{\text{err}})\big) \\
\rho_1 &\sim \mathcal{N}\big(\bar{\rho}_1 \coloneqq (n-m)(1-p_{\text{err}}), {\sigma_\rho}^2\big),
\end{aligned}$$

where the detector picks a Lebesgue measurable function $f : \mathbb{R} \to [0,1]$ and the attacker a pair of distributions $(\omega_0,\omega_1) \in \mathcal{A}$ with support on the interval $I = [0,m]$; and one *discrete game* defined by the similar criteria (43) but with four binomial distributions

$$\begin{aligned}
\tilde{\sigma}_0 &\sim \text{Binom}(n, p_{\text{err}}), & \tilde{\sigma}_1 &\sim \text{Binom}(n, 1-p_{\text{err}}), \\
\tilde{\rho}_0 &\sim \text{Binom}\,\big(n-m, p_{\text{err}}\big), & \tilde{\rho}_1 &\sim \text{Binom}\,\big(n-m, 1-p_{\text{err}}\big).
\end{aligned}$$

where the detector picks a piecewise constant function $f : \mathbb{R} \to [0, 1]$ of the form

$$f(\bar{z}) = \sum_{k=-\infty}^{\infty} f_k \Delta(\bar{z} - k), \qquad \Delta(s) \coloneqq \begin{cases} 1 & s \in [\eta, \eta + 1) \\ 0 & \text{otherwise} \end{cases} \tag{51}$$

with $\eta \in [-1/2, 0]$ and $f_k \in \{0, 1\}$, $\forall k$ and the attacker a pair of (discrete) distributions $(\omega_0, \omega_1) \in \mathcal{A}$ with support on the discrete set $\{0, 1, \ldots, m\}$. In the discrete game, the random variable $\bar{Z}$ observed by the detector is always an integer between 0 and n and therefore we can indeed restrict the estimation policies f to be of the form (51). Note that the value of f in (51) on the integers $\{0, 1, \ldots, n\}$ is exactly the same regardless of the choice of η in the interval $\eta \in [-1/2, 0]$. We shall select the value for η shortly to make sure that we can use Lemma 3.

In view of Corollary 1, the continuous game has a saddle-point (38)–(39) provided that (32) holds and that (33) has a single zero. The former is implied by (36) and the latter holds in view of the assumption that (37) has a single zero z^*. Moreover, by selecting η equal to the distance between z^* and its closest integer, we can make sure that the step-like saddle point f^* in (35) is of the form (51). To apply Lemma 2, it remains to show that the criteria J and $\tilde{J}$ of the continuous and discrete games, respectively, do not differ by more than ϵ for discrete distributions ω_0, ω_1 with support on the set $\{0, 1, \ldots, m\}$ and piecewise constant functions f of the form (51). This result is provided by Lemma 3 with ϵ defined by (46) and

$$\begin{aligned}
c_0 &\coloneqq \sup_{i \in \mathbb{Z}, f_k \in \{0,1\}} \left| \int_{\mathbb{R}} f(\bar{y} + i) \big(\sigma_0(d\bar{y}) - \tilde{\sigma}_0(d\bar{y}) \big) \right|, \\
c_1 &\coloneqq \sup_{i \in \mathbb{Z}, f_k \in \{0,1\}} \left| \int_{\mathbb{R}} f(\bar{y} + i) \big(\sigma_1(d\bar{y}) - \tilde{\sigma}_1(d\bar{y}) \big) \right|, \\
d_0 &\coloneqq \sup_{i \in \mathbb{Z}, f_k \in \{0,1\}} \left| \int_{\mathbb{R}} f(\bar{y}) \big(\rho_0(d\bar{y}) - \tilde{\rho}_0(d\bar{y}) \big) \right|, \\
d_1 &\coloneqq \sup_{i \in \mathbb{Z}, f_k \in \{0,1\}} \left| \int_{\mathbb{R}} f(\bar{y}) \big(\rho_1(d\bar{y}) - \tilde{\rho}_1(d\bar{y}) \big) \right|,
\end{aligned}$$

for f as in (51). To compute c_0, we expand the integral as follows

$$c_0 = \sup_{i \in \mathbb{Z}, f_k \in \{0,1\}} \left| \sum_{k=-\infty}^{\infty} f_{k+i} \int_{k+\eta}^{k+\eta+1} \big(\sigma_0(d\bar{y}) - \tilde{\sigma}_0(d\bar{y}) \big) \right|$$

and note that the largest (positive) value of the summation will take place when the f_{k+i} are equal to 1 when $\int_{k+\eta}^{k+\eta+1} \big(\sigma_0(d\bar{y}) - \tilde{\sigma}_0(d\bar{y}) \big) > 0$ and equal to 0 otherwise. Similarly, the most negative

value for the summation will take place when the f_{k+i} are equal to 1 when $\int_{k+\eta}^{k+\eta+1}\big(\sigma_0(d\bar{y}) - \tilde{\sigma}_0(d\bar{y})\big) < 0$ and equal to 0 otherwise. This leads to

$$\begin{aligned} c_0 = \sup_{i\in\mathbb{Z}, f_k\in\{0,1\}} \Bigg| & \sum_{k=-\infty}^{-1} f_{k+i} \int_{k+\eta}^{k+\eta+1} \sigma_0(d\bar{y}) + \sum_{k=n+1}^{\infty} f_{k+i} \int_{k+\eta}^{k+\eta+1} \sigma_0(d\bar{y}) \\ & \sum_{k=0}^{n} f_{k+i}\bigg(\int_{k+\eta}^{k+\eta+1} \sigma_0(d\bar{y}) - \binom{n}{k} p_{\text{err}}^k (1-p_{\text{err}})^{n-k}\bigg)\Bigg| \\ = \max\Bigg\{ & \int_{-\infty}^{\eta} \sigma_0(d\bar{y}) + \int_{n+1+\eta}^{\infty} \sigma_0(d\bar{y}) \;\; \sum_{k=0}^{n} H\bigg(\int_{k+\eta}^{k+\eta+1} \sigma_0(d\bar{y}) - \binom{n}{k} p_{\text{err}}^k (1-p_{\text{err}})^{n-k}\bigg), \\ & - \sum_{k=0}^{n} H\bigg(\binom{n}{k} p_{\text{err}}^k (1-p_{\text{err}})^{n-k} - \int_{k+\eta}^{k+\eta+1} \sigma_0(d\bar{y})\bigg)\Bigg\} = \mathcal{E}(p_{\text{err}}, n). \end{aligned}$$

The formulas for c_1, d_0, d_1 can be similarly derived. ■

V. Conclusions and future work

This paper introduced novel game-theoretic approaches to estimate a binary random variable based on a vector of binary sensor measurements that may have been corrupted by an attacker, also known as the Byzantine problem. The problem is formulated as a zero-sum partial information game in which a detector attempts to minimize the probability of an estimation error and an attacker attempts to maximize this probability. We provide two complementary solutions to this problem: the first builds upon policy domination to provide an optimal estimator that is valid when the number of sensors n does not exceed $2/p_{\text{err}}$ and the second provides a suboptimal estimator that is, at most, ϵ-away from the optimal with ϵ converging to zero as fas as $1/\sqrt{n}$. The two approaches are complementary in that they cover both a small and large number of sensors. The results presented are limited to the estimation of a static random variable so the key question for future work is an extension of this work to the estimation of states of a dynamical system, either in the form of a finite state machine, a Markov chain, or a stochastic differential or difference equation.

Appendix

Proof of Lemma 1. By the law of total probability, we can expand

$$\mathrm{P}_{\mu,\delta}(\hat{X} \neq X) = (1-p)\sum_{k=0}^{n} \mathrm{P}_{\mu}\left(\hat{X} = 1 \middle| \sum_{i=1}^{n} Z_i = k\right) \mathrm{P}_{\delta}\left(\sum_{i=1}^{n} Z_i = k \middle| X = 0\right)$$

$$+ p \sum_{k=0}^{n} \mathrm{P}_{\mu}\left(\hat{X}=0 \middle| \sum_{i=1}^{n} Z_i = k\right) \mathrm{P}_{\mu}\left(\sum_{i=1}^{n} Z_i = k \middle| X=1\right)$$

$$= (1-p) \sum_{k=0}^{n} \mu(k)\, \mathrm{P}_{\delta}\left(\sum_{i=1}^{n} Z_i = k \middle| X=0\right) + p \sum_{k=0}^{n} \big(1-\mu(k)\big)\, \mathrm{P}_{\delta}\left(\sum_{i=1}^{n} Z_i = k \middle| X=1\right), \quad (52)$$

where we used the facts that

$$\mathrm{P}_{\mu\delta}\left(\hat{X}=1 \middle| \sum_{i=1}^{n} Z_i = k\right) = \mu(k),$$

$$\mathrm{P}_{\mu\delta}\left(\hat{X}=0 \middle| \sum_{i=1}^{n} Z_i = k\right) = 1-\mu(k).$$

We now proceed to compute the conditional probabilities in (52), which can also be expanded as follows:

$$\mathrm{P}_{\delta}\left(\sum_{i=1}^{n} Z_i = k \middle| X\right) = \mathrm{P}_{\delta}\left(\sum_{i=1}^{n} Z_i = k \middle| X, \mathcal{E}_{\text{attack}}\right) p_{\text{attack}} + \mathrm{P}\left(\sum_{i=1}^{n} Y_i = k \middle| X\right)(1-p_{\text{attack}}),$$

where $\mathcal{E}_{\text{attack}}$ denotes the events that the attacker manipulated measurements. When no measurements have been manipulated, the random variable of interest $\sum_{i=1}^{n} Y_i$ has a binomial distribution and we simply have that

$$\mathrm{P}\left(\sum_{i=1}^{n} Y_i = k \middle| X\right) = \binom{n}{k} \begin{cases} p_{\text{err}}^{k}(1-p_{\text{err}})^{n-k} & X=0, \\ (1-p_{\text{err}})^{k} p_{\text{err}}^{n-k} & X=1. \end{cases} \quad (53)$$

Otherwise, since $\delta(X)$ sets to 0 and to 1 a number of sensors equal to $\delta(X)$ and to $m-\delta(X)$, respectively, we have that

$$\mathrm{P}_{\delta}\left(\sum_{i=1}^{n} Z_i = k | X, \mathcal{E}_{\text{attack}}\right) = \begin{cases} f(X), & m-\delta(X) \leqslant k \leqslant n-\delta(X) \\ 0, & \text{otherwise} \end{cases} \quad (54)$$

where

$$f(X) = \binom{n-m}{k-m+\delta(X)} p_{\text{err}}^{k-m+\delta(X)} (1-p_{\text{err}})^{n-k-\delta(X)}.$$

Equation (5) follows from (53), (54), and (52). ■

Proof of Lemma 2. Because of (42) and the fact that (u^*, d^*) is a saddle-point for J, we conclude that

$$\tilde{J}(u^*, d^*) \;\leqslant\; \epsilon + J(u^*, d^*) \;\leqslant\; \epsilon + J(u, d^*) \;\leqslant\; 2\epsilon + \tilde{J}(u, d^*), \qquad \forall u \in \tilde{\mathcal{U}}.$$

and also that

$$\tilde{J}(u^*,d^*) \geqslant J(u^*,d^*) - \epsilon \geqslant J(u^*,d) - \epsilon \geqslant \tilde{J}(u^*,d) - 2\epsilon, \quad \forall d \in \tilde{\mathcal{D}}$$

from which we conclude that (u^*,d^*) is indeed a 2ϵ saddle-point for $\tilde{J}$. ■

Proof of Proposition 1. We use an induction argument on k to prove (17): The basis of induction $k=1$ follows from the equality:

$$\binom{n}{1}\Big(p_{\rm err}(1-p_{\rm err})^{n-1} - p_{\rm err}^{n-1}(1-p_{\rm err})\Big) = n\Big(p_{\rm err}(1-p_{\rm err})^{n-1} - p_{\rm err}^{n-1}(1-p_{\rm err})\Big).$$

To prove the induction step we assume now that (17) holds for some integer k, such that $1 \leqslant k \leqslant n-2$, and evaluate the left hand side of (17) with k replaced by $k+1$:

$$\binom{n}{k+1}\Big(p_{\rm err}^{k+1}(1-p_{\rm err})^{n-k-1} - p_{\rm err}^{n-k-1}(1-p_{\rm err})^{k+1}\Big)$$
$$= \binom{n}{k}\left(\frac{n-k}{k+1}\frac{p_{\rm err}}{1-p_{\rm err}}p_{\rm err}^{k}(1-p_{\rm err})^{n-k} - \frac{n-k}{k+1}\frac{1-p_{\rm err}}{p_{\rm err}}p_{\rm err}^{n-k}(1-p_{\rm err})^{k}\right)$$

Since

$$\frac{n-k}{k+1}\frac{p_{\rm err}}{1-p_{\rm err}} \leqslant 1 \;\Leftrightarrow\; p_{\rm err} \leqslant \frac{k+1}{n+1} \;\Leftarrow\; p_{\rm err} \leqslant \frac{2}{n+1}$$

and

$$\frac{n-k}{k+1}\frac{1-p_{\rm err}}{p_{\rm err}} \geqslant 1 \;\Leftrightarrow\; p_{\rm err} \leqslant \frac{n-k}{n+1} \;\Leftarrow\; p_{\rm err} \leqslant \frac{2}{n+1},$$

we conclude that

$$\binom{n}{k+1}\Big(p_{\rm err}^{k+1}(1-p_{\rm err})^{n-k-1} - p_{\rm err}^{n-k-1}(1-p_{\rm err})^{k+1}\Big)$$
$$\leqslant \binom{n}{k}\big(p_{\rm err}^{k}(1-p_{\rm err})^{n-k} - p_{\rm err}^{n-k}(1-p_{\rm err})^{k}\big) \leqslant n\big(p_{\rm err}(1-p_{\rm err})^{n-1} - p_{\rm err}^{n-1}(1-p_{\rm err})\big),$$

where the last inequality follows from the assumption that (17) holds for every integer k such that $1 \leqslant k \leqslant n-1$. This concludes the induction argument. ■

Proposition 2: Consider a function

$$g(z) \coloneqq e^{-a_2z^2+b_1z+b_0} - e^{-a_2z^2+c_1z+c_0} + e^{-\bar{a}_2z^2+\bar{b}_1z+\bar{b}_0} - e^{-\bar{a}_2z^2+\bar{c}_1z+\bar{c}_0}, \tag{55}$$

with

$$a_2, \bar{a}_2 > 0, \quad \max\{b_1,\bar{b}_1\} < \min\{c_1,\bar{c}_1\} \tag{56}$$

There exists a constant $\delta > 0$ such that when $|a_2 - \bar{a}_2| < \delta$ the function $g(z)$ has a single zero at $z = z^*$, is negative before z^* and positive after z^*. □

Proof of Proposition 2. First note that we can re-write $g(z)$ as

$$g(z) = e^{-a_2 z^2 + \beta z} \bar{g}(z)$$

with

$$\bar{g}(z) \coloneqq \bar{g}_1(z) + e^{(a_2 - \bar{a}_2) z^2} \bar{g}_2(z) \tag{57}$$

$$\bar{g}_1(z) \coloneqq e^{(b_1 - \beta) z + b_0} - e^{(c_1 - \beta) z + c_0} \tag{58}$$

$$\bar{g}_2(z) \coloneqq e^{(\bar{b}_1 - \beta) z + \bar{b}_0} - e^{(\bar{c}_1 - \beta) z + \bar{c}_0}, \tag{59}$$

and β such that

$$b_1 < \beta < c_1, \quad \bar{b}_1 < \beta < \bar{c}_1. \tag{60}$$

Such β exists because of (56). In view of this, to prove that $g(z)$ has a single zero, it suffices to show that $\bar{g}(z)$ has a single zero. First we note that, because of (60), we have that

$$\lim_{\bar{z} \to -\infty} \bar{g}(\bar{z}) = +\infty, \qquad \lim_{\bar{z} \to \infty} \bar{g}(\bar{z}) = -\infty,$$

and therefore, by continuity, $\bar{g}(z)$ must have at least one zero. To show that this function has a single zero, we prove that it is strictly monotonically decreasing. Taking derivatives of the two functions in (58)–(59) with respect to z, we obtain

$$\bar{g}_1'(z) = (b_1 - \beta) e^{(b_1 - \beta) z + b_0} - (c_1 - \beta) e^{(c_1 - \beta) z + c_0}$$

$$\bar{g}_2'(z) = (\bar{b}_1 - \beta) e^{(\bar{b}_1 - \beta) z + \bar{b}_0} - (\bar{c}_1 - \beta) e^{(\bar{c}_1 - \beta) z + \bar{c}_0},$$

which in view of (56) shows that $\bar{g}_1$ and $\bar{g}_2$ are strictly monotone decreasing. Since $\bar{g}_2$ is strictly monotone decreasing and

$$\lim_{\bar{z} \to -\infty} \bar{g}_2(\bar{z}) = +\infty, \qquad \lim_{\bar{z} \to \infty} \bar{g}_2(\bar{z}) = -\infty,$$

this function has a single zero at some point z^*. To show that $e^{(a_2 - \bar{a}_2) z^2} \bar{g}_2(z)$ is also strictly monotone decreasing we consider three cases

1) for $z \leqslant 0$ and $z \leqslant z^*$, $e^{(a_2 - \bar{a}_2) z^2} \bar{g}_2(z)$ is the product of two positive monotone decreasing functions and therefore is monotone decreasing

2) for $z \geqslant 0$ and $z \geqslant z^*$, $e^{(a_2-\bar{a}_2)z^2}\bar{g}_2(z)$ is the product of a positive monotone increasing function by a negative monotone decreasing function and therefore is monotone decreasing
3) between 0 and z^*, the derivative of $e^{(a_2-\bar{a}_2)z^2}\bar{g}_2(z)$ satisfies

$$e^{(a_2-\bar{a}_2)z^2}\Big(\big(2(a_2-\bar{a}_2)z+(\bar{b}_1-\beta)\big)e^{(\bar{b}_1-\beta)z+\bar{b}_0}-\big(2(a_2-\bar{a}_2)z+(\bar{c}_1-\beta)\big)e^{(\bar{c}_1-\beta)z+\bar{c}_0}\Big)$$

and we can always pick $a_2-\bar{a}_2$ sufficiently small so that the sign of $\big(2(a_2-\bar{a}_2)z+(\bar{b}_1-\beta)\big)$ and $(2(a_2-\bar{a}_2)z+(\bar{c}_1-\beta))$ does not change in this interval. ■

REFERENCES

[1] T. Alpcan, and T. Basar, "A Game Theoretic Analysis of Intrusion Detection in Access Control Systems," in *Proc. 43rd IEEE Conference on Decision and Control (CDC)*, December, vol. 2, pp.1568-1573, 2004.

[2] J. S. Baras, "Security and Trust for Wireless Autonomic Networks: System and Control Methods," *European Journal of Control*, vol. 13, no. (2-3), pp. 105-133, 2007.

[3] T. Basar and G. J. Olsder, *Dynamic Noncooperative Game Theory*, 2nd ed. Philadelphia, PA:SIAM, 1999.

[4] S. Becker, J. Seibert, D. Zage, C. Nita-Rotaru, R. State,"Applying game theory to analyze attacks and defenses in virtual coordinate systems," in *Proc. 2011 IEEE/IFIP 41st International Conference on Dependable Systems and Networks (DSN)*, pp.133-144, June 2011

[5] E. Byres and J. Lowe, "The myths and facts behind cyber security risks for industrial control systems," *VDE Congress*, 2004.

[6] A. A. Cardenas, S. Amin, and S. Sastry, "Research challenges for the security of control systems," *HOTSEC08: Proceedings of the 3rd conference on Hot topics in security*, pp. 1-6, 2008.

[7] Z. E. Fuchs, P. P. Khargonekar, "Games, deception and Jones' Lemma," *Proc. American Control Conference (ACC)*, pp. 4532-4537, 2011.

[8] J. P. Hespanha, Y. S. Ateskan, H. H. Kzlocak, "Deception in non cooperative games with partial information," *Proc. of the 2nd DARPA-JFACC Symp. on Advances in Enterprise Control*, 2000.

[9] O. Kosut and L. Tong, "Capacity of cooperative fusion in the presence of Byzantine sensors," *Proc. 44th Annu. Allerton Conf. Commun., Contr. Comput.*, Monticello, IL, Sep. 2729, 2006.

[10] L. Lamport, R. Shostak, and M. Pease, "The Byzantine generals problem," *ACM Transactions on Programming Languages and Systems*, vol. 4, no. 3, 1982.

[11] A. S. Matveev, A. V. Savkin, *Estimation and Control over Communication Networks*, Birkhauser Boston, 2008.

[12] W. McEneaney, "Some classes of imperfect information finite state space stochastic games with finite-dimensional solutions," *Applied Mathematics and Optimization*, Vol. 50, no. 2, pp. 87-118, 2004.

[13] Y. Mo, J. Hespanha, B. Sinopoli, "Robust Detection in the Presence of Integrity Attacks," In *Proc. of the 2012 Amer. Contr. Conf.*, pp. 3541-3546, 2012.

[14] V. Srinivasan, V. Nuggehalli, C-F. Chiasserini, and R. R. Rao, "An Analytical Approach to the Study of Cooperation in Wireless Ad Hoc Networks," *IEEE Transactions on Communications*, March, vol. 4, No. 2, pp.722-733, 2005.

[15] A. Teixeira, S. Amin, H. Sandberg, K. H. Johansson, and S. S. Sastry, "Cyber-security analysis of state estimators in electric power systems," *IEEE CDC*, pp. 5991-5998, 2010.

[16] S. Tijs, *Introduction to Game Theory*, Hindustan Book Agency, India, 2003.

[17] M. Tubaishat and S. Madria, "Sensor Networks: an Overview," *IEEE Potentials*, vol. 22, no. 2, 20-23, April 2003.

[18] A. Urpi, M. Bonuccelli, and S. Giordano, "Modelling Cooperation in Mobile Ad Hoc Networks: A formal description of selfishness," *WiOpt'03 Workshop: Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, 2003.

[19] K. G. Vamvoudakis, J. P. Hespanha, B. Sinopoli, Y. Mo, "Adversarial Detection as a Zero-Sum Game," *Proc. IEEE Conference on Decision and Control*, pp. 7133-7138, Maui, Hawaii, 2012.

[20] J. Weiss, *Protecting Industrial Control Systems from Electronic Threats*, second edition, Momentum Press, Philadelphia, 2010.