Passive Fault-tolerant Estimation under Strategic Adversarial Bias

Serkan Sarıtaş, György Dán and Henrik Sandberg

***Abstract*— This paper is concerned with the problem of fault-tolerant estimation in cyber-physical systems. In cyber-physical systems, such as critical infrastructures, networked embedded sensors are widely used for monitoring and can be exploited by an adversary to deceive the control center by modifying measured values. The deception is modeled as a bias; i.e., there is a misalignment between the objective functions of the control center and the adversarial sensor. Different from previous studies, a Stackelberg equilibrium of a cheap talk setup is adapted to the attacker-defender game setting for the first time. That is, the defender (control center), as a receiver, is the leader, and the attacker (adversarial sensor), as a transmitter, is the follower. The equilibrium strategies and the associated costs are characterized for uniformly distributed variables and quadratic objective functions, and an analysis on the uniqueness of the equilibrium is provided. It is shown that the attacker and defender costs at the equilibrium are increasing with the bias and decreasing with the number of quantization levels. Our results surprisingly show that, under certain conditions, the attacker prefers a public bias rather than a private one.**

I. INTRODUCTION

Modern critical infrastructures (CI), such as electric power systems, gas and water distribution, are prominent examples of cyber-phisycal systems (CPS). They depend on large-scale industrial control systems for safe and efficient operation, often referred to as Supervisory Control and Data Acquisition (SCADA) systems. SCADA systems collect measurement data from remote terminal units and deliver the measurement data to a master station located at a control center. At the control center the data are typically fed into a state estimator (SE), which provides an accurate estimate of the system's state despite noisy or faulty measurement data collected by the SCADA system [1], and control actions are taken based on the estimated state. As an example, enabled by recent developments in measurement and communication technology, the operation of electric power systems increasingly relies on high frequency real-time monitoring and SE [2].

Ensuring the security of SE is thus fundamental for CIs, but it is a significant challenge due to the emergence of cyber-physical attacks, as those could often remain undetected [3]. One approach for ensuring proper operation despite attacks is to treat attacks as faults, and to design control systems capable of maintaining a desired level of performance despite faults. Such a control technique is referred to as fault tolerant control (FTC). FTC strategies can be classified into two categories: passive and active methods [4]. In a passive approach, the control system is designed so that it maintains the designed performance under healthy as well as faulty system situations that have been considered at the design stage, without any change in the control law. Therefore, the controller, which remains fixed during the entire system operation (i.e., does not require a reconfiguration), is able to maintain the stability of the system with an acceptable degradation of its performance, whatever the system situation (healthy or faulty). In contrast, an active approach reacts to each fault situation immediately by properly adapting the controller design.

Going beyond random faults typically considered in FTC, an attacker with knowledge of the system would arguably induce a worst case fault for a given FTC strategy. This calls for a modeling approach that considers the strategic interaction between the attacker and FTC. To address this, in this paper we consider a fault-tolerant estimation problem in the presence of a strategic adversary that is capable of manipulating sensory data, and her objective is to introduce adversarial bias in estimation. Such adversarial bias could result in, e.g., the manipulation of generation dispatch and of power markets [5], but it could also be used to hide that the power system is in an unsafe state and could eventually lead to cascading failures.

In our model the control center (i.e., the defender) aims at designing a passive fault-tolerant estimator for the underlying state/physical variable, while the objective of the adversary is to introduce a bias in the estimate. As a consequence the objectives of the control center and of the adversary are misaligned, akin to the problem of cheap talk signaling studied in the economics literature initiated by Crawford and Sobel [6]. Contrary to existing works on cheap talk [6]–[14] in our work we formulate the problem as a Stackelberg game, where the receiver is the leader and the transmitter is the follower. We argue that this modeling assumption is necessary in a security context in order to capture the worst case attack [15], and is consistent with Kerckhoffs's principle and Shannon's assumption that *"the enemy knows the system"* [16]. To the best of our knowledge, we are the first to propose a cheap-talk Stackelberg game formulation of the attacker-defender problem in the context of fault-tolerant and secure estimation.

This work was supported in part by the Swedish Research Council (grant 2016-00861), and the Swedish Civil Contingencies Agency (MSB) through the CERCES project.

S. Sarıtaş is with the Division of Information Science and Engineering, KTH Royal Institute of Technology, SE-10044, Stockholm, Sweden. saritas@kth.se

G. Dán is with the Division of Network and Systems Engineering, KTH Royal Institute of Technology, SE-10044, Stockholm, Sweden. gyuri@kth.se

H. Sandberg is with the Division of Decision and Control Systems, KTH Royal Institute of Technology, SE-10044, Stockholm, Sweden. hsan@kth.se

A. Related Work

Early works on cheap talk investigated Nash equilibrium strategies, i.e., the attacker and the defender announce their strategies simultaneously, and found that surprisingly, under some technical conditions on the objective functions of the players, the cheap talk problem only admits equilibrium strategies that are essentially quantized [6]. This is in significant contrast with the case where the objective functions are aligned. Subsequent works considered extensions of the Nash equilibrium of cheap talk, e.g., [8], [12] with similar observations.

Another line of works considered the Stackelberg equilibrium of cheap talk [7]–[14]. In these works the transmitter (i.e., the attacker) is the leader of the game, whereas the receiver (i.e., the defender) is the follower. This is, however, inconsistent with the usual modeling framework in the security domain [17], where the defender is the leader and acts first by committing to a strategy, while the attacker is the follower and chooses how and where to attack after observing the defender's choice [18]. Our work thus complements existing work on cheap talk by considering the Stackelberg equilibrium for attacker-defender scenarios common in the security domain, where the defender is the leader.

A brief overview of on passive, active, and hybrid fault-tolerant estimation and control can be found in [19]. Several passive FTC methods have been proposed in the literature, mainly based on robust theory; however, due to their inflexibility and low performance, the active and hybrid methods are more popular for fault-tolerant estimation [4]. Among them, the (active) fault-tolerant estimation problem is addressed by means of the reconfiguration of the sensor network, a strategy by which only the subset of healthy sensors is used in [20], [21]. Furthermore, related to ours are works on secure estimation for CPS, where the CPS is typically modeled as a linear time-invariant (LTI) system in which multiple sensors measure the same physical variable, and some sensors might be under attack. By utilizing combinatorial-type approaches (which are active FTC methods), a characterization of the number of attacked sensors that can be tolerated was investigated under different assumptions [22]–[24]. Different from these studies, our work focuses on the game theoretic aspect of estimation using passive methods, rather than the combinatorial problem of choosing sensors to be attacked or protected.

B. Contributions

The main contributions of this work are as follows:

(i) The fault-tolerant and secure estimation problem under adversarial bias is formulated as a cheap talk problem, and the Stackelberg equilibrium of the corresponding cheap talk setup is adapted to the attacker-defender game setting for the first time.

(ii) Equilibrium strategies and the associated costs are characterized for uniformly distributed variables and quadratic objective functions, and the effects of the (public and private) bias and the number of quantization levels on the equilibrium costs are investigated.

(iii) It is shown that the control center is always better off with a public bias. However, contrary to expectations, the adversarial sensor also prefers a public bias, rather than a private one, for a small bias.

The rest of the paper is organized as follows. We present the problem formulation in Section II, and analyze the Stackelberg equilibria with public and private bias in Section III and Section IV, respectively. In Section V we provide numerical examples, and Section VI concludes the paper.

II. Problem Formulation

The cheap talk problem can be formulated as the following 2-player game. An informed player (adversarial sensor) knows the value of the $\mathbb{X}$-valued state/source X (e.g., voltage level) and transmits the $\mathbb{M}$-valued message M to the other player (control center), who takes her $\mathbb{X}$-valued optimal action U (e.g., secure estimation in the current formulation) upon receiving the message M. The strategies of the adversarial sensor and the control center are assumed to be deterministic; i.e., $M = f(X)$ and $U = g(M) = g(f(X))$. Let $c^a(x,u)$ and $c^d(x,u)$ denote the (Borel measurable) objective functions of the attacker (adversarial sensor) and the defender (control center), respectively, when the action u is taken for the corresponding state x. Then, for the given strategies, the attacker's induced expected cost is $J^a(f,g) = \mathbb{E}_X\left[c^a(X,U)\right]$, whereas the defender's induced expected cost is $J^d(f,g) = \mathbb{E}_X\left[c^d(X,U)\right]$.

As opposed to the original cheap talk formulation in [6], we investigate the Stackelberg equilibrium: the defender (control center) is the leader, while the attacker (adversarial sensor) is the follower[1]. Consequently, the defender's strategy is known to the attacker, and this fact is known to the leader. Note that assuming the opposite, i.e., that the attacker does not know the defender's strategy, would be equivalent to assuming that the defender's strategy is a cryptographic secret, which is unrealistic considering that it is an algorithm. Since the attacker chooses her strategy based on the strategy of the defender, the strategy of the attacker can be represented by $M = f_g(X)$, where f_g represents the dependency of the attacker strategy on the defender strategy. Then, a pair of strategies $(f^*_{g^*}, g^*)$ is a Stackelberg equilibrium [15] if

$$J^d(f^*_{g^*}, g^*) \leq J^d(f^*_g, g) \quad \forall g \in \Gamma^d\,,$$

where f^*_g satisfies

$$J^a(f^*_g, g) \leq J^a(f_g, g) \quad \forall f_g \in \Gamma^a\,,$$

where Γ^a and Γ^d are the sets of all deterministic (and Borel measurable) functions from $\mathbb{X}$ to $\mathbb{M}$ and from $\mathbb{M}$ to $\mathbb{X}$, respectively. In the Stackelberg game, the leader cannot backtrack on her commitment (which validates a passive fault-tolerant approach; i.e., the estimator/controller does not require a reconfiguration during the entire system operation),

[1]Throughout the manuscript, the terms control center, defender, and leader are used interchangeably. Similarly, the terms adversarial sensor, attacker, and follower are used interchangeably.

but has a leadership role since she can manipulate the follower by anticipating the follower's actions.

We assume real valued variables and quadratic objective functions; i.e., $\mathbb{X} = \mathbb{M} = \mathbb{R}$, $c^a(x,u) = (x-u-b)^2$ and $c^d(x,u) = (x-u)^2$, where b denotes a bias that is strategically introduced to the system by an adversarial sensor.

The state X to be estimated by the control center is assumed to be a uniform random variable (r.v.) (see Section II-A), and without loss of generality, it can be assumed to be a standard uniform r.v.; i.e., $X \sim \mathcal{U}[0,1]$. The bias b can be private or public. When the bias is private, i.e., known to the adversarial sensor only, the control center's prior is that b is uniformly distributed on $[-1,1]$; i.e., $b \sim \mathcal{U}[-1,1]$, which is independent[2] of X. For the public bias case, b is available to both players; i.e., both players know each other's objective functions. Considering the case of the public bias provides an insight about the relation between the defender's strategy and the attacker's objective, which can be helpful to make comparisons with the private bias case. In some scenarios, the adversarial sensor may also have a well-known incentive (economical, for instance) to add a certain bias.

A. Motivational Example

Consider a smart electric power grid[3]: the control center has to ensure that the voltage level at some nodes (clients/households) is in the allowed range by collecting voltage data from the sensors located around the nodes. The nodes may prefer to report an incorrect or a biased measurement within the allowed range. For example, even though the voltage level rises, individual households who produce energy from photovoltaics (PV) may report a lower voltage in order to sell more energy to the system (notice that the bias can be modeled as public for this case since the control center may be aware of such incentives). Assuming that volt/VAR control (VVC) is integrated to the smart power distribution systems, voltage levels can be preserved within acceptable ranges (95 to 105 % of nominal) [26]. Since the nominal voltage levels are between $220-240$ Volts, it can be assumed that each voltage is equally likely for the analysis. Regarding the problem formulation above, X is the actual voltage level, M is the transmitted (biased) measurement, and U is the control center's estimate.

III. EQUILIBRIA WITH A PUBLIC BIAS

In this section, the Stackelberg equilibrium is analyzed for a fixed and public bias; i.e., b is known by both the attacker and the defender. For an announced defender strategy g, the goal of the attacker is to minimize her cost $J^a(f_g,g) = \mathbb{E}_X\left[(X-g(f_g(X))-b)^2\right]$. On the other hand, due to the leadership role, by anticipating the optimal attacker strategy f_g^*, the defender aims to minimize $J^d(f_g^*,g) = \mathbb{E}_X\left[\left(X-g(f_g^*(X))\right)^2\right]$. Before presenting the technical results, we provide the following observation which is valid for any type of source distribution with a public bias.

Observation 3.1: For any given invertible defender strategy $g(M)$, the optimal attacker strategy is $f_g^*(X) = g^{-1}(X-b)$, which implies $J^a(f_g^*,g)=0$. In this case, the defender cost becomes $J^d(f_g^*,g)=b^2$.

The observation above shows that if the defender uses an invertible strategy, the incurred cost to the defender is always b^2. We thus investigate whether a lower cost can be achieved with other types of strategies. As the simplest one, suppose that $g(M)=c$, where c is a real constant. The optimal choice of c is given in the following observation, and is valid for any type of source distribution with public bias.

Observation 3.2: For a given defender strategy $g(M)=c$, the incurred cost to the attacker is $J^a(f_g,g) = \mathbb{E}_X\left[(X-c-b)^2\right]$, which is independent of the strategy of the attacker. Therefore, the cost to be minimized by the defender is $J^d(f_g,g) = \mathbb{E}_X\left[(X-c)^2\right]$, which results in $c^* = \mathbb{E}_X[X]$ as the optimal defender strategy. In this case, the costs of the attacker and the defender at the equilibrium are $J^a(f_{g^*},g^*) = \mathrm{Var}(X)+b^2$ and $J^d(f_{g^*},g^*) = \mathrm{Var}(X)$, respectively, where $\mathrm{Var}(\cdot)$ denotes the variance of a r.v..

Notice that under the equilibrium described in Observation 3.2, the attacker's actions are irrelevant and the defender takes an action based only on her priors. Such an equilibrium is called as a non-informative (babbling) equilibrium [6]. Further, as it can be seen, as long as $\mathrm{Var}(X) < b^2$, the defender prefers a non-informative equilibrium over any equilibria with an invertible strategy. At this point, it is worth to analyze the equilibrium costs when the defender has more than one action to choose from.

Observation 3.3: For a uniform source $X \sim \mathcal{U}[0,1]$ and a given defender with two actions u_1 and u_2 such that $0 \le u_1 \le u_2 \le 1$, the attacker minimizes her cost $J^a(f_g,g) = \mathbb{E}_X\left[(X-g(f_g(X))-b)^2\right]$ by minimizing $(x-g(f_g(x))-b)^2$ for every possible value of X with a proper choice of strategy $f_g(X)$. The attacker prefers $g(f_g(x))=u_1$ if $(x-u_1-b)^2 < (x-u_2-b)^2 \Rightarrow x < \frac{u_1+u_2}{2}+b$, and $g(f_g(x))=u_2$ if $x \ge \frac{u_1+u_2}{2}+b$. Note that, the attacker is indifferent between $g(f_g(x))=u_1$ and $g(f_g(x))=u_2$ when $x=\frac{u_1+u_2}{2}+b$. Further, if $\frac{u_1+u_2}{2}+b \le 0$, the best response of the attacker satisfies $g(f_g^*(x))=u_2$ for every $x \in [0,1]$; hence, the optimal defender action is $u_2^* = \mathbb{E}_X[X] = \frac{1}{2}$. Similarly, if $\frac{u_1+u_2}{2}+b \ge 1$, the optimal attacker and defender actions are $g(f_g^*(x))=u_1$ and $u_1^*=\frac{1}{2}$, respectively. Thus, the optimal attacker strategy satisfies

$$g(f_g^*(x)) = \begin{cases} u_2 & \frac{u_1+u_2}{2}+b \le 0\,,\ \forall x \in [0,1] \\ u_1 & 0 \le x < \frac{u_1+u_2}{2}+b \\ u_2 & \frac{u_1+u_2}{2}+b \le x < 1 \\ u_1 & \frac{u_1+u_2}{2}+b \ge 1\,,\ \forall x \in [0,1] \end{cases}. \quad (1)$$

The observation above shows that, unless $0 < \frac{u_1+u_2}{2}+b <$

[2]For this case, even though the biased measurement $X-b$ is out of the interval $[0,1]$, the estimate of the control center is always on the interval $[0,1]$ (the voltage levels are assumed to be preserved within an acceptable range, see Section II-A). The extension to the correlated bias and source is left as a future study.

[3]Under a smart grid scenario with a quite similar setting but slightly modified objective functions, the Nash equilibria of a signaling game between a consumer and an electricity aggregator are investigated in [25].

1, the defender with two actions is essentially equivalent to the defender with a single action (see Observation 3.2). Otherwise; i.e., when $0 < \frac{u_1+u_2}{2} + b < 1$ holds, the interaction between the attacker and the defender can be represented as a quantization game[4] in which the reconstruction values (i.e., quantization levels) are determined by the leader (defender) and the boundaries are determined by the follower (attacker). Namely, as shown in Fig. 1, the defender selects the optimal quantization levels u_1 and u_2 by anticipating the boundaries (that are determined by the attacker via the nearest neighbor condition), and thus, the voltage levels are quantized at the equilibrium.

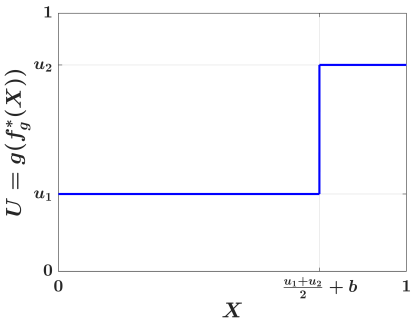

Fig. 1: The best response of the attacker is represented for the defender actions (i.e., quantization levels) $u_1 = 0.3$ and $u_2 = 0.8$ when $b = 0.2$. Notice that the optimal boundary is chosen as $\frac{u_1+u_2}{2} + b = 0.75$ by the attacker.

The following theorem investigates the equilibrium (i.e., the optimal quantizer for the defender) with two defender actions (i.e., quantization levels). The proof is in Appendix A.

Theorem 3.1: For a uniformly distributed source $X \sim \mathcal{U}[0,1]$ and publicly known bias b, suppose that the defender has two quantization levels such that $0 \le u_1 \le u_2 \le 1$. Then, the equilibrium is characterized as follows:

	u_1^*	u_2^*	$J^a(f_{g^*}^*, g^*)$	$J^d(f_{g^*}^*, g^*)$
$\|b\| < \frac{1}{2}$	$\frac{1}{4} + b^2$	$\frac{3}{4} - b^2$	$3\left(b^2 + \frac{1}{12}\right)^2$	$-\left(b^2 - \frac{1}{4}\right)^2 + \frac{1}{12}$
$\|b\| \ge \frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$b^2 + \frac{1}{12}$	$\frac{1}{12}$

Theorem 3.1 describes the equilibrium with two quantization levels. The following theorem characterizes equilibria with $N > 2$ quantization levels. The proof is in Appendix B.

Theorem 3.2: For a uniformly distributed source $X \sim \mathcal{U}[0,1]$ and publicly known bias b, suppose that the defender has $N > 2$ quantization levels such that $u_{[1:N]} \triangleq u_1, u_2, \ldots, u_N$ with $0 \le u_1 \le u_2 \le \ldots \le u_N \le 1$. Let $t \triangleq N-1$, $A \triangleq \sqrt{4b^2(t^2-1)+1}$, and $\Delta = \frac{t-A}{t^2-1}$.

(i) If $|b| < \frac{1}{2}$, the optimal quantization levels are $u_1^* = \frac{-1+At}{2(t^2-1)}$ and $u_i^* = u_1^* + (i-1)\Delta$ for $i = 1, 2, \ldots, N$, and the corresponding attacker and defender costs are $J^a(f_{g^*}^*, g^*) = \frac{4tA^3 - 3A^2 + t^2 - 6tA + 4}{12(t^2-1)^2}$ and $J^d(f_{g^*}^*, g^*) = \frac{-2tA^3 + 3t^2A^2 - 2t^2 + 1}{12(t^2-1)^2}$, respectively.

(ii) Otherwise; i.e., $|b| \ge \frac{1}{2}$, there exist only babbling equilibria under which $u_i^* = \frac{1}{2}$ for $i = 1, 2, \ldots, N$, $J^a(f_{g^*}^*, g^*) = b^2 + \frac{1}{12}$ and $J^d(f_{g^*}^*, g^*) = \frac{1}{12}$.

Theorem 3.2 characterizes the non-babbling equilibrium for $|b| < \frac{1}{2}$. Therefore, there are two extreme conditions to investigate:

Corollary 3.1: (i) For $b = 0$, the costs of the attacker and the defender become the same. Then, for any N, it follows that $A = 1$, $u_1^* = \frac{1}{2N}$, $\Delta = \frac{1}{N}$, and $J^d(f_{g^*}^*, g^*) = \frac{1}{12N^2}$. Notice that the attacker partitions the source as $[0, \frac{1}{N}], \ldots, (\frac{t}{N}, 1]$, which is a uniform quantizer.

(ii) For $|b| = \frac{1}{2}$ case, for any $N > 2$, it follows that $A = N-1$, $\Delta = 0$, $u_1^* = \ldots = u_N^* = \frac{1}{2}$, and $J^d(f_{g^*}^*, g^*) = \frac{1}{12}$. As it can be seen, N different quantization levels converge into the one, resulting in a babbling equilibrium.

Theorem 3.2 characterizes the equilibrium depending on the number of quantization levels N and bias b, and allows us to analyze their impact on the attacker and the defender:

Corollary 3.2: The costs of the attacker and the defender are increasing functions of $|b|$ and decreasing functions of N. Thus, both the attacker and the defender prefer a smaller $|b|$ and a larger number of quantization levels N.

Remark 3.1: (i) Regarding Observation 3.1, there are infinitely many invertible strategies of the defender resulting in the same cost for both players, thus the corresponding equilibria are outcome equivalent.

(ii) Regarding Observation 3.2, there are infinitely many possible strategies of the attacker resulting in the same cost for both players, thus the corresponding equilibria are outcome equivalent.

(iii) Regarding Theorem 3.1 and Theorem 3.2, there are infinitely many strategy pairs $(f_{g^*}^*, g^*)$ resulting in the unique optimal quantizer levels u_i^* and unique equilibrium costs $J^a(f_{g^*}^*, g^*)$ and $J^d(f_{g^*}^*, g^*)$ for fixed N and b, thus the corresponding equilibria are outcome equivalent for fixed N and b.

Remark 3.2: (i) For a fixed number of quantization levels N, there is a critical value $\tilde{b}$ such that as long as $|b| < \tilde{b}$, the defender prefers an invertible strategy over a quantized one with N quantization levels. This follows from the fact that as $b \to 0$, the quantized defender cost converges to $\lim_{b\to 0} J^d(f_{g^*}^*, g^*) = \frac{1}{12N^2}$ (i.e., a uniform quantizer); whereas, the cost of the defender with an invertible strategy converges to 0: $\lim_{b\to 0} J^d(f_{g^*}^*, g^*) = b^2 \to 0$.

(ii) For a fixed bias b, there is a critical value $\widetilde{N}$ such that the defender prefers a strategy with $N > \widetilde{N}$ quantization levels over an invertible one. This follows from the fact that as $N \to \infty$, $A \to 2|b|(N-1)$ and $J^d(f_{g^*}^*, g^*) \to b^2(1 - \frac{4|b|}{3}) < b^2$.

(iii) If we let $t = N-1 \to \infty$ and $b \to 0$ simultaneously with $|bt| = C \gg 1$, where C is a fixed constant, since $A \to 2|b|t$, we have

$$\lim_{\substack{t\to\infty,\, b\to 0 \\ tb = C \gg 1}} J^d(f_{g^*}^*, g^*) \approx b^2\,, \qquad \lim_{\substack{t\to\infty,\, b\to 0 \\ tb = C \gg 1}} J^a(f_{g^*}^*, g^*) \to 0\,.$$

This shows that when $t \to \infty$ and $b \to 0$ simultaneously with $|bt| = C \gg 1$, the costs converge to those in Observation 3.1.

[4]Note that the quantization strategy of the attacker is not an arbitrary choice among non-invertible strategies, it is the best response for a given defender strategy with countably many actions.

(iv) $J^a(f_g^*, g^*) = J^d(f_g^*, g^*) + b^2(4u_1^* - 1)$ holds by Theorem 3.2. Thus, the attacker cost is greater than or equal to the defender cost iff $u_1^* \geq \frac{1}{4}$. Note that, for $N = 2$, since $u_1^* = \frac{1}{4} + b^2$, this condition is always satisfied. However, this does not necessarily hold for $N > 2$. Since $u_1^* = \frac{-1+t\sqrt{4b^2(t^2-1)+1}}{2(t^2-1)}$, $J^a(f_g^*, g^*) < J^d(f_g^*, g^*)$ holds if $16b^2t^2 - t^2 + 1 < 0$. In particular, for a fixed b, if $t^2 > \frac{1}{1-16b^2}$, then the attacker has a lower cost. Similarly, for a fixed t, if $b^2 < \frac{t^2-1}{16t^2}$, then the attacker has a lower cost.

IV. EQUILIBRIA WITH A PRIVATE BIAS

In this section, the Stackelberg equilibrium is analyzed for a random and private bias; i.e., the prior of the defender is the distribution of the bias, which is assumed to be $b \sim \mathcal{U}[-1, 1]$. Due to the random bias, the defender does not know the attacker's objective function perfectly, and the defender cost includes the expectation with respect to both X and b; i.e., $J^d(f_g^*, g) = \mathbb{E}_{X,b}\left[\left(X - g(f_g^*(X))\right)^2\right]$. On the other hand, since the realization of b is available to the attacker, her cost remains $J^a(f_g, g) = \mathbb{E}_X\left[\left(X - g(f_g(X)) - b\right)^2\right]$. Before presenting the technical results, we provide the following observations, which are valid for any type of source distribution with a private bias.

Observation 4.1: For any given invertible defender strategy $g(M)$, the optimal attacker strategy is $f_g^*(X) = g^{-1}(X - b)$, which results in $J^a(f_g^*, g) = 0$. In this case, the defender cost is $J^d(f_g^*, g) = \mathbb{E}_b[b^2]$.

Observation 4.2: For a defender with a single action (i.e., quantization level), the optimal strategy is $g^*(M) = c^* = \mathbb{E}_X[X]$, which results in $J^a(f_{g^*}, g^*) = \mathbb{E}_b[\mathrm{Var}(X) + b^2] = \mathrm{Var}(X) + \mathbb{E}_b[b^2]$ and $J^d(f_{g^*}, g^*) = \mathrm{Var}(X)$, respectively.

The equilibrium with a private bias when the defender has two quantization levels is analyzed below. The proof is in Appendix C.

Theorem 4.1: For a uniformly distributed source $X \sim \mathcal{U}[0, 1]$ and bias $b \sim \mathcal{U}[-1, 1]$ (which is independent of X), whose realization is available only to the attacker, suppose that the defender has two quantization levels such that $0 \leq u_1 \leq u_2 \leq 1$. Then, the optimal quantization levels are $u_1^* = \frac{5}{12}$ and $u_2^* = \frac{7}{12}$, and, as a function of the realization of b, the corresponding equilibrium costs are as follows:

	$\boldsymbol{\lvert b\rvert < \frac{1}{2}}$	$\boldsymbol{\lvert b\rvert \geq \frac{1}{2}}$
$\boldsymbol{J_b^a(f_{g^*}^*, g^*)}$	$\frac{5b^2}{6} + \frac{7}{144}$	$b^2 - \frac{\lvert b\rvert}{6} + \frac{13}{144}$
$\boldsymbol{J_b^d(f_{g^*}^*, g^*)}$	$\frac{b^2}{6} + \frac{7}{144}$	$\frac{13}{144}$

The analysis with two quantization levels in Theorem 4.1 can be extended to N quantization levels as follows. The proof is in Appendix D.

Theorem 4.2: For a uniformly distributed source $X \sim \mathcal{U}[0, 1]$ and bias $b \sim \mathcal{U}[-1, 1]$ (which is independent of X), whose realization is available only to the attacker, suppose that the defender has $N > 2$ quantization levels such that $u_{[1:N]} \triangleq u_1, u_2, \ldots, u_N$ with $0 \leq u_1 \leq u_2 \leq \ldots \leq u_N \leq 1$. Let $t \triangleq N - 1$, $A \triangleq \sqrt{\frac{2t^2+1}{3}}$, and $\Delta = \frac{t-A}{t^2-1}$. Then the optimal quantization levels can be characterized as $u_1^* = \frac{-1+At}{2(t^2-1)}$ and $u_i^* = u_1^* + (i-1)\Delta$ for $i = 1, 2, \ldots, N$.

Remark 4.1: Similar to Remark 3.1, the outcome equivalence of the equilibria can be established for Observation 4.1, Observation 4.2, Theorem 4.1, and Theorem 4.2.

V. NUMERICAL EXAMPLE AND DISCUSSION

In this section, we present an example, Fig. 2, to illustrate the analytical results and make comparisons between the scenarios with a public and a private bias. As it can be seen from Fig. 2, for a public bias, both the defender and the attacker prefer the equilibrium with greater number of quantization levels (see Corollary 3.2). The same behavior can be observed for a private bias (considering the realizations of b), too. Furthermore, as shown in Remark 3.2.(iv), for a public bias, the attacker has a lower cost than the defender for $b^2 < \frac{2^2-1}{16\times 2^2} \Rightarrow |b| < 0.2165$ when the defender has three quantization levels; whereas, the attacker has always a higher cost when the defender has two quantization levels. Again, the same behavior can be observed for a private bias, too.

If the defender knows the bias, intuitively, she optimizes her cost better than in the case when she knows only the distribution of the bias; i.e., the defender cost with a private bias is higher than with a public one. Indeed, for the two quantization levels case and $|b| \leq \frac{1}{2}$, $-\left(b^2 - \frac{1}{4}\right)^2 + \frac{1}{12} \geq \frac{b^2}{6} + \frac{7}{144}$ is always true, with equality holding only when $b^2 = \frac{1}{6}$, as depicted in Fig. 2. On the other hand, for the two quantization levels case, unexpectedly, the attacker prefers a public bias, rather than a private one, for smaller bias values; i.e., she has a higher cost for a private bias. Accordingly, the attacker prefers a public bias when $3\left(b^2 + \frac{1}{12}\right)^2 \leq \frac{5b^2}{6} + \frac{7}{144} \Rightarrow b^2 \leq \frac{1}{6}$.

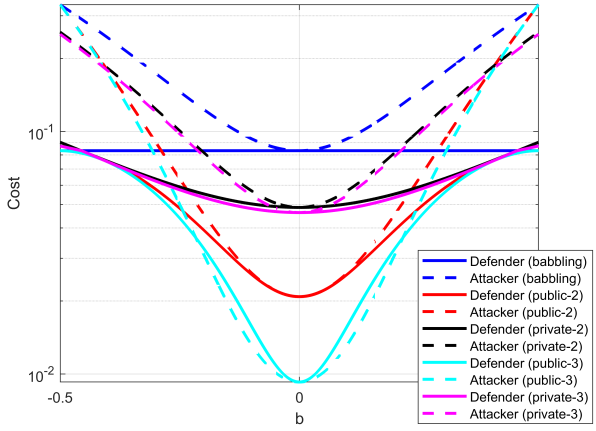

Fig. 2: Comparison of the attacker and defender costs with respect to the bias under different scenarios (public and private bias) and number of quantization levels (1, 2, and 3).

VI. CONCLUSIONS AND FUTURE WORKS

A fault-tolerant and secure estimation problem under strategic adversarial bias was modeled as a cheap talk game

between the control center and an adversarial sensor. Stackelberg equilibria of the corresponding game were investigated under the public and private bias assumptions, the conditions for outcome equivalence and uniqueness of the equilibria were established, and the equilibrium strategies and the associated costs were characterized. It was shown that the cost of the attacker and of the defender is an increasing function of bias and a decreasing function of number of quantization levels. Our results surprisingly show that, under certain conditions, the attacker prefers a public bias rather than a private one.

For the public bias case, when the defender decides an invertible strategy, the cost of the defender is b^2 whereas the attacker achieves zero cost (see Observation 3.1). On the other hand, a non-invertible (e.g., quantized) defender strategy allows to share the cost between the players. Even though the total cost of the players increases, the attacker cost becomes more than zero, which can be desirable by the defender, and the defender even decreases her own cost from b^2 by using quantized strategies (see Remark 3.2.(ii)).

Our model has many possible interesting extensions. Of particular interest are the case when the source and the bias are correlated, the case of more general (e.g., unbounded) source and bias distributions, such as Gaussian, and the case of dynamic (multi-stage) interaction. Furthermore, the setup can be extended to multiple control centers (i.e., hierarchical or distributed SE) and/or multiple sensors (that can be either honest or adversarial).

APPENDIX

A. Proof of Theorem 3.1

Due to Observation 3.3 and (1), for the given quantization levels u_1 and u_2 with $0 < \frac{u_1+u_2}{2} + b < 1$, the corresponding defender cost is

$$J^d(f_g^*, g) = \int_0^{\frac{u_1+u_2}{2}+b} (x-u_1)^2 \mathrm{d}x + \int_{\frac{u_1+u_2}{2}+b}^1 (x-u_2)^2 \mathrm{d}x$$
$$= \frac{1}{12} + \left(u_2 - \frac{1}{2}\right)^2 + (u_2-u_1)\left(b^2 - \frac{(u_2+u_1)^2}{4}\right). \tag{2}$$

The defender aims to minimize (2) by selecting the optimal actions u_1 and u_2 under the constraints $0 \le u_1 \le u_2 \le 1$ and $0 \le \frac{u_1+u_2}{2} + b \le 1$. Since the 2×2 Hessian matrix $\mathbb{H} = \left[\frac{\partial^2 J^d(f_g^*, g)}{\partial u_i \partial u_j}\right]$ of (2) is positive semi-definite, $J^d(f_g^*, g)$ in (2) is a convex function of u_1 and u_2. Further, the inequality constraints are convex (actually they are affine). Thus, the optimization problem of the defender is convex [27]. The corresponding Lagrangian function is expressed as

$$\mathcal{L}(u_1, u_2, \lambda, \mu, \nu, \gamma, \zeta) = \frac{1}{12} + \left(u_2 - \frac{1}{2}\right)^2 + b^2(u_2-u_1)$$
$$- \frac{(u_2-u_1)(u_2+u_1)^2}{4} - \lambda u_1 + \mu(u_1-u_2) + \nu(u_2-1)$$
$$- \gamma\left(\frac{u_1+u_2}{2} + b\right) + \zeta\left(\frac{u_1+u_2}{2} + b - 1\right),$$

and the dual function is given by

$$h(\lambda, \mu, \nu, \gamma, \zeta) \triangleq \inf_{u_1, u_2} \quad \mathcal{L}(u_1, u_2, \lambda, \mu, \nu, \gamma, \zeta),$$

and the Lagrangian dual problem is defined as

$$\min_{\lambda,\mu,\nu,\gamma,\zeta} h(\lambda, \mu, \nu, \gamma, \zeta) \text{ s.t. } \lambda, \mu, \nu, \gamma, \zeta \ge 0.$$

Since the optimization problem is convex, the duality gap between the solutions of the primal and the dual problem is zero. Then, the Karush-Kuhn-Tucker (KKT) conditions (stationarity, primal feasibility, dual feasibility, and complementary slackness) can be obtained as (3)-(6), respectively.

$$\frac{\partial \mathcal{L}(u_1, u_2, \lambda, \mu, \nu, \gamma, \zeta)}{\partial u_i} = 0 \text{ for } i = 1, 2, \tag{3}$$
$$-u_1 \le 0,\ u_1 - u_2 \le 0,\ u_2 - 1 \le 0,$$
$$-\left(\frac{u_1+u_2}{2} + b\right) \le 0,\ \frac{u_1+u_2}{2} + b - 1 \le 0, \tag{4}$$
$$\lambda \ge 0,\ \mu \ge 0,\ \nu \ge 0,\ \gamma \ge 0,\ \zeta \ge 0, \tag{5}$$
$$\lambda u_1 = 0,\ \mu(u_1-u_2) = 0,\ \nu(u_2-1) = 0,$$
$$\gamma\left(\frac{u_1+u_2}{2} + b\right) = 0,\ \zeta\left(\frac{u_1+u_2}{2} + b - 1\right) = 0. \tag{6}$$

By utilizing the conditions above, $0 < u_1 < u_2 < 1$ and $\lambda = \nu = \gamma = \zeta = 0$ are obtained, and (3) reduces to $u_1^2 = (u_2-1)^2 \Rightarrow u_1 = 1 - u_2$. Then, the stationarity condition in (3) becomes $\frac{\partial \mathcal{L}(u_1,u_2,\lambda,\mu,\nu,\gamma,\zeta)}{\partial u_1} = 0 \Rightarrow u_1 = b^2 + \frac{1}{4} - \mu$ and $u_2 = \frac{3}{4} - b^2 + \mu$. Due to (6), $\mu \ne 0$ implies $u_1 = u_2 \Rightarrow \mu = b^2 - \frac{1}{4}$. Thus, if $b^2 \ge \frac{1}{4}$, it follows that $u_1^* = u_2^* = \frac{1}{2}$, which is a babbling equilibrium as described in Observation 3.2, and the corresponding attacker and defender costs become $J^a(f_{g^*}^*, g^*) = b^2 + \frac{1}{12}$ and $J^d(f_{g^*}^*, g^*) = \frac{1}{12}$, respectively.

On the other hand, $\mu = b^2 - \frac{1}{4}$ contradicts with (5) if $b^2 < \frac{1}{4}$. Therefore, if $b^2 < \frac{1}{4}$, we can conclude that $\mu = 0$, and the corresponding optimal quantization levels become $u_1^* = b^2 + \frac{1}{4}$ and $u_2^* = \frac{3}{4} - b^2$. Then, the corresponding defender cost in (2) becomes $J^d(f_{g^*}^*, g^*) = -\left(b^2 - \frac{1}{4}\right)^2 + \frac{1}{12}$. Similarly, the corresponding attacker cost can be calculated. ■

B. Proof of Theorem 3.2

For the given quantization levels of defender $u_{[1:N]}$, the best response of the attacker is characterized as $g(f_g(x)) = \arg\min_{\tilde{u}=u_{[1:N]}} (x - \tilde{u} - b)^2$, which is equivalent to

$$g(f_g^*(x)) = \begin{cases} u_1 & 0 \le x \le \frac{u_1+u_2}{2} + b \\ u_i & \frac{u_{i-1}+u_i}{2} + b < x \le \frac{u_i+u_{i+1}}{2} + b \\ u_N & \frac{u_{N-1}+u_N}{2} + b < x < 1 \end{cases}, \tag{7}$$

for $i = 2, 3, \ldots, N-1$. Note that

- if $\frac{u_1+u_2}{2} + b \le 0$, then the optimal attacker response satisfies $g(f_g^*(x)) = u_2$ for $0 \le x \le \frac{u_2+u_3}{2} + b$, and the quantization level $g(f_g(x)) = u_1$ cannot be utilized.
- if $\frac{u_{N-1}+u_N}{2} + b \ge 1$, then the optimal attacker satisfies $g(f_g^*(x)) = u_{N-1}$ for $\frac{u_{N-2}+u_{N-1}}{2} + b < x \le 1$, and the quantization level $g(f_g(x)) = u_N$ cannot be utilized.

Therefore, unless $0 < \frac{u_1+u_2}{2} + b$ and $\frac{u_{N-1}+u_N}{2} + b < 1$, the N-level quantization setup reduces to the $(N-1)$-level

quantization setup. Then, for $0 < \frac{u_1+u_2}{2}+b$ and $\frac{u_{N-1}+u_N}{2}+b<1$, the corresponding defender cost is

$$\begin{aligned} J^d(f_g^*,g) &= \sum_{i=2}^{N-1}\left(\int_{\frac{u_{i-1}+u_i}{2}+b}^{\frac{u_i+u_{i+1}}{2}+b}(x-u_i)^2\mathrm{d}x\right) \\ &+\int_0^{\frac{u_1+u_2}{2}+b}(x-u_1)^2\mathrm{d}x+\int_{\frac{u_{N-1}+u_N}{2}+b}^1(x-u_N)^2\mathrm{d}x \\ &=\frac{1}{12}+\left(u_N-\frac{1}{2}\right)^2+b^2(u_N-u_1) \\ &\quad -\frac{1}{4}\sum_{i=2}^{N}(u_i-u_{i-1})(u_i+u_{i-1})^2\,. \end{aligned} \tag{8}$$

The defender aims to minimize (8) by selecting the optimal actions $u_{[1:N]}$ under the constraints $0\le u_1\le u_2\le\ldots\le u_N\le 1$, $0\le\frac{u_1+u_2}{2}+b$ and $\frac{u_{N-1}+u_N}{2}+b\le 1$. Since the $N\times N$ Hessian matrix $\mathbb{H}=\left[\frac{\partial^2 J^d(f_g^*,g)}{\partial u_i\partial u_j}\right]$ of (8) is symmetric and diagonally dominant matrix with real non-negative diagonal entries; its eigenvalues are real, and by Gershgorin's circle theorem, all of its eigenvalues are non-negative. Thus, $\mathbb{H}$ is positive semi-definite, which implies that $J^d(f_g^*,g)$ in (8) is a convex function of $u_{[1:N]}$. Further, the inequality constraint functions are convex. Thus, the optimization problem of the defender is convex [27]. The corresponding Lagrangian function is expressed as

$$\begin{aligned} \mathcal{L}\left(u_{[1:N]},\lambda,\mu_{[1:N-1]},\nu,\gamma,\zeta\right) &= \frac{1}{12}+\left(u_N-\frac{1}{2}\right)^2 \\ &+b^2(u_N-u_1)-\frac{1}{4}\sum_{i=2}^{N}(u_i-u_{i-1})(u_i+u_{i-1})^2 \\ &-\lambda u_1+\left(\sum_{i=1}^{N-1}\mu_i(u_i-u_{i+1})\right)+\nu(u_N-1) \\ &-\gamma\left(\frac{u_1+u_2}{2}+b\right)+\zeta\left(\frac{u_{N-1}+u_N}{2}+b-1\right), \end{aligned}$$

and the dual function and the Lagrangian dual problem can be defined in a similar way to that in the proof of Theorem 3.1. Since the optimization problem is convex, the duality gap between the solutions of the primal and the dual problem is zero. Then, by letting $\Delta_i=u_{i+1}-u_i$, the KKT conditions (stationarity, primal feasibility, dual feasibility, and complementary slackness) can be obtained as (9)-(12), respectively.

$$\frac{\partial\mathcal{L}\left(u_{[1:N]},\lambda,\mu_{[1:N-1]},\nu,\gamma,\zeta\right)}{\partial u_i}=0 \text{ for } i=1,\ldots,N\,, \tag{9}$$

$$\begin{aligned} &-u_1\le 0\,,\ -\Delta_i\le 0 \text{ for } i=1,\ldots,N-1\,,\ u_N-1\le 0\,, \\ &-\left(\frac{u_1+u_2}{2}+b\right)\le 0\,,\ \frac{u_{N-1}+u_N}{2}+b-1\le 0\,, \end{aligned} \tag{10}$$

$$\lambda\ge 0\,,\ \mu_i\ge 0 \text{ for } i=1,2,\ldots,N-1\,,\ \nu\ge 0\,, \tag{11}$$

$$\begin{aligned} &\lambda u_1=0\,,\ \mu_i\Delta_i=0 \text{ for } i=1,\ldots,N-1\,,\ \nu(u_N-1)=0\,, \\ &\gamma\left(\frac{u_1+u_2}{2}+b\right)=0\,,\ \zeta\left(\frac{u_{N-1}+u_N}{2}+b-1\right)=0\,. \end{aligned} \tag{12}$$

By utilizing the conditions above, $0<u_1\le u_2\le\ldots\le u_N<1$ and $\lambda=\nu=\gamma=\zeta=0$ are obtained, and (9) reduces to $u_1^2=(u_N-1)^2\Rightarrow u_1=1-u_N$. Then, the stationarity condition in (9) becomes $\frac{\partial\mathcal{L}\left(u_{[1:N]},\lambda,\mu_{[1:N-1]},\nu,\gamma,\zeta\right)}{\partial u_i}=0\Rightarrow\frac{\Delta_{i-1}^2}{4}-\mu_{i-1}=\frac{\Delta_i^2}{4}-\mu_i$ for $i=2,3,\ldots,N-1$. Hence,

$$\frac{\Delta_1^2}{4}-\mu_1=\frac{\Delta_2^2}{4}-\mu_2=\ldots=\frac{\Delta_N^2}{4}-\mu_N \tag{13}$$

holds. In (13), if any $\Delta_j=0$, then $\frac{\Delta_j^2}{4}-\mu_j\le 0$ by (12), which requires $\frac{\Delta_i^2}{4}-\mu_i\le 0$ for all $i=2,3,\ldots,N-1$. Thus, it must hold that $\Delta_1=\ldots=\Delta_{N-1}=0$, which implies $u_1=\ldots=u_N$ and $\mu_1=\ldots=\mu_{N-1}$. Since $u_1+u_N=1$, we have $u_1=\ldots=u_N=\frac{1}{2}$. Further, the stationarity condition in (9) becomes $\frac{\partial\mathcal{L}\left(u_{[1:N]},\lambda,\mu_{[1:N-1]},\nu,\gamma,\zeta\right)}{\partial u_1}=0\Rightarrow\mu_1=\ldots=\mu_{N-1}=b^2-\frac{1}{4}$. Thus, if $b^2\ge\frac{1}{4}$, it follows that $u_1^*=\ldots=u_N^*=\frac{1}{2}$, which is a babbling equilibrium as described in Observation 3.2, and the corresponding attacker and defender costs become $J^a(f_{g^*}^*,g^*)=b^2+\frac{1}{12}$ and $J^d(f_{g^*}^*,g^*)=\frac{1}{12}$, respectively.

On the other hand, $\mu=b^2-\frac{1}{4}$ contradicts with (11) if $b^2<\frac{1}{4}$. Therefore, if $b^2<\frac{1}{4}$, $\Delta_i>0$ must hold for $i=1,2,\ldots,N-1$, which implies $\mu_1=\ldots=\mu_{N-1}=0\Rightarrow\Delta_1=\ldots=\Delta_{N-1}\triangleq\Delta$. Thus, the relation between the optimal quantization levels becomes $u_i=u_1+(i-1)\Delta$ for $i=1,2,\ldots,N$. Since $u_1+u_N=1$, the interval between the quantization levels is $\Delta=\frac{1-2u_1}{N-1}$. Then, by letting $t\triangleq N-1$ and $A\triangleq\sqrt{4b^2(t^2-1)+1}$, the stationarity condition with respect to u_1 in (9); i.e., $\frac{\partial\mathcal{L}\left(u_{[1:N]},\lambda,\mu_{[1:N-1]},\nu,\gamma,\zeta\right)}{\partial u_1}=0$, implies

$$\left(1-\frac{1}{t^2}\right)(u_1)^2+\frac{1}{t^2}u_1-\left(b^2+\frac{1}{4t^2}\right)=0\,. \tag{14}$$

Since $u_1>0$, the only possible solution of the quadratic equation in (14) is $u_1^*=\frac{-1+At}{2(t^2-1)}$. Then, $\Delta=\frac{1-2u_1^*}{N-1}=\frac{t-A}{t^2-1}$ and $u_i^*=u_1^*+(i-1)\Delta$ are obtained. Further, the condition $0\le\frac{u_1^*+u_2^*}{2}+b$ reduces to $b\ge-\frac{1}{2}$, and $\frac{u_{N-1}^*+u_N^*}{2}+b\le 1$ reduces to $b\le\frac{1}{2}$. Therefore, the $b^2<\frac{1}{4}$ assumption holds. Then, the optimal defender cost in (8) becomes $J^d(f_{g^*}^*,g^*)=\frac{-2tA^3+3t^2A^2-2t^2+1}{12(t^2-1)^2}$. Similarly, the corresponding attacker cost can be calculated. ■

C. Proof of Theorem 4.1

Since the attacker has a perfect observation of b, for every realization of b, her optimal strategy satisfies (1). Then, the corresponding average defender cost becomes

$$\begin{aligned} J^d(f_g^*,g) &= \mathsf{Pr}\left(-1\le b\le-\frac{u_1+u_2}{2}\right)\int_0^1(x-u_2)^2\mathrm{d}x \\ &+\mathsf{Pr}\left(-\frac{u_1+u_2}{2}<b<1-\frac{u_1+u_2}{2}\right) \\ &\quad\times\mathbb{E}_b\left[\int_0^{\frac{u_1+u_2}{2}+b}(x-u_1)^2\mathrm{d}x\right. \\ &\quad\left.+\int_{\frac{u_1+u_2}{2}+b}^1(x-u_2)^2\mathrm{d}x\,\middle|\,-\frac{u_1+u_2}{2}<b<1-\frac{u_1+u_2}{2}\right] \end{aligned}$$

$$+\,\mathsf{Pr}\left(1-\frac{u_1+u_2}{2}\leq b\leq 1\right)\int_0^1 (x-u_1)^2\mathrm{d}x$$
$$=\frac{1}{12}+\left(u_2-\frac{1}{2}\right)^2+(u_2-u_1)\left(\frac{1}{6}-\frac{(u_2+u_1)^2}{4}\right). \tag{15}$$

Due to the resemblance between (2) and (15), the corresponding optimal quantization levels are $u_1^*=\frac{5}{12}$ and $u_2^*=\frac{7}{12}$, and the optimal average defender cost is $J^d(f_{g^*}^*,g^*)=\frac{11}{144}$. Then, the corresponding defender and attacker costs can be derived as a function of the realization of b. ■

D. Proof of Theorem 4.2

If $\frac{u_{N-1}+u_N}{2}+b\leq 0$, then the best response of the attacker will always end up with the quantization level u_N; i.e., $g(f_g^*(x))=u_N\,\forall x\in[0,1]$. However, if $\frac{u_{N-2}+u_{N-1}}{2}+b\leq 0$, then her best response will result in 2 different quantization levels u_N and u_{N-1}; in particular, $g(f_g^*(x))=u_N$ if $x\in[0,\frac{u_{N-1}+u_N}{2}+b]$, and $g(f_g^*(x))=u_{N-1}$ if $x\in(\frac{u_{N-1}+u_N}{2}+b,1]$. Note that it is impossible to have $\frac{u_i+u_{i+1}}{2}+b\leq 0$ and $\frac{u_j+u_{j+1}}{2}+b\geq 1$ simultaneously for any $i,j=1,2,\ldots,N$, unless $b=0$ (which has zero measure). Then, proceeding similarly, the optimal attacker strategy can be characterized by utilizing Table I.

TABLE I: The optimal attacker strategy and available quantization levels (for $k=2,3,\ldots,N-1$)

Condition	Available Quantization Levels
$-1\leq b\leq -\frac{u_{N-1}+u_N}{2}$	u_N
$-\frac{u_k+u_{k+1}}{2}<b\leq -\frac{u_{k-1}+u_k}{2}$	$u_N,u_{N-1},\ldots,u_k$
$-\frac{u_1+u_2}{2}<b\leq 1-\frac{u_{N-1}+u_N}{2}$	$u_N,u_{N-1},\ldots,u_1$
$1-\frac{u_k+u_{k+1}}{2}<b\leq 1-\frac{u_{k-1}+u_k}{2}$	$u_k,u_{k-1},\ldots,u_1$
$1-\frac{u_1+u_2}{2}<b\leq 1$	u_1

The attacker determines her best response by considering the quantization levels available to the defender, which is indicated in Table I; e.g., if $1-\frac{u_k+u_{k+1}}{2}\leq b\leq 1-\frac{u_{k-1}+u_k}{2}$, then the attacker chooses boundaries between k quantization levels whose reconstruction values are $u_1,u_2,\ldots,u_k$, similar to that in (7). Then, by utilizing (8), the corresponding average defender cost can be calculated as

$$J^d(f_{g^*}^*,g^*)=\frac{1}{12}+\left(u_N-\frac{1}{2}\right)^2+\frac{u_N-u_1}{6}-\frac{1}{4}\sum_{k=2}^{N}(u_k-u_{k-1})(u_k+u_{k-1})^2. \tag{16}$$

Due to the resemblance between (8) and (16), the corresponding optimal quantization levels are $u_1^*=\frac{-1+At}{2(t^2-1)}$, $\Delta=\frac{t-A}{t^2-1}$, and $u_i^*=u_1^*+(i-1)\Delta$ for $i=1,2,\ldots,N$, with $t=N-1$ and $A=\sqrt{\frac{2t^2+1}{3}}$. ■

REFERENCES

[1] A. Monticelli, "Electric power system state estimation," *Proceedings of the IEEE*, vol. 88, no. 2, pp. 262–282, Feb. 2000.

[2] A. Ipakchi and F. Albuyeh, "Grid of the future," *IEEE Power and Energy Magazine*, vol. 7, no. 2, pp. 52–62, Mar. 2009.

[3] J. Giraldo, E. Sarkar, A. A. Cardenas, M. Maniatakos, and M. Kantarcioglu, "Security and privacy in cyber-physical systems: A survey of surveys," *IEEE Design Test*, vol. 34, no. 4, pp. 7–17, Aug. 2017.

[4] M. Blanke, M. Kinnaert, J. Lunze, and M. Staroswiecki, *Reconfigurability Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 343–387.

[5] A. Teixeira, H. Sandberg, G. Dán, and K. H. Johansson, "Optimal power flow: Closing the loop over corrupted data," in *American Control Conference (ACC)*, June 2012, pp. 3534–3540.

[6] V. P. Crawford and J. Sobel, "Strategic information transmission," *Econometrica*, vol. 50, pp. 1431–1451, 1982.

[7] E. Kamenica and M. Gentzkow, "Bayesian persuasion," *American Economic Review*, vol. 101, no. 6, pp. 2590–2615, Oct. 2011.

[8] S. Sarıtaş, S. Yüksel, and S. Gezici, "Quadratic multi-dimensional signaling games and affine equilibria," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 605–619, Feb. 2017.

[9] F. Farokhi, A. M. H. Teixeira, and C. Langbort, "Estimation with strategic sensors," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 724–739, Feb. 2017.

[10] E. Akyol, C. Langbort, and T. Başar, "Information-theoretic approach to strategic communication as a hierarchical game," *Proceedings of the IEEE*, vol. 105, no. 2, pp. 205–218, Feb. 2017.

[11] M. L. Treust and T. Tomala, "Persuasion with limited communication capacity," *Journal of Economic Theory*, vol. 184, p. 104940, 2019.

[12] S. Sarıtaş, S. Yüksel, and S. Gezici, "Dynamic signaling games with quadratic criteria under Nash and Stackelberg equilibria," *Automatica*, vol. 115, p. 108883, May 2020.

[13] M. O. Sayin, E. Akyol, and T. Başar, "Hierarchical multistage Gaussian signaling games in noncooperative communication and control systems," *Automatica*, vol. 107, pp. 9–20, 2019.

[14] M. O. Sayin and T. Başar, "Dynamic information disclosure for deception," in *57th IEEE Conference on Decision and Control (CDC)*, Dec. 2018, pp. 1110–1117.

[15] T. Başar and G. J. Olsder, *Dynamic Noncooperative Game Theory*. Philadelphia, PA: SIAM Classics in Applied Mathematics, 1999.

[16] C. E. Shannon, "Communication theory of secrecy systems," *The Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, Oct. 1949.

[17] G. Brown, M. Carlyle, J. Salmern, and K. Wood, "Defending critical infrastructure," *Interfaces*, vol. 36, no. 6, pp. 530–544, 2006.

[18] M. Jain, J. Pita, M. Tambe, F. Ordóñez, P. Paruchuri, and S. Kraus, "Bayesian Stackelberg games and their application for security at Los Angeles International Airport," *SIGecom Exchanges*, vol. 7, no. 2, pp. 10:1–10:3, June 2008.

[19] J. Jiang, "Fault-tolerant control systems–An introductory overview," *Acta Automatica Sinica*, vol. 31, no. 1, p. 161, 2005.

[20] L. H. Mutuel and J. L. Speyer, "Fault-tolerant estimation," in *American Control Conference (ACC)*, June 2000, pp. 3718–3722.

[21] M. Staroswiecki, G. Hoblos, and A. Aitouche, "Sensor network design for fault tolerant estimation," *International Journal of Adaptive Control and Signal Processing*, vol. 18, no. 1, pp. 55–72, 2004.

[22] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1454–1467, June 2014.

[23] Y. Mo and B. Sinopoli, "Secure estimation in the presence of integrity attacks," *IEEE Transactions on Automatic Control*, vol. 60, no. 4, pp. 1145–1151, Apr. 2015.

[24] O. Vuković, K. C. Sou, G. Dán, and H. Sandberg, "Network-aware mitigation of data integrity attacks on power system state estimation," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 6, pp. 1108–1118, July 2012.

[25] B. Larrousse, O. Beaude, and S. Lasaulce, "Crawford-Sobel meet Lloyd-Max on the grid," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 6127–6131.

[26] A. Teixeira, G. Dán, H. Sandberg, R. Berthier, R. B. Bobba, and A. Valdes, "Security of smart distribution grids: Data integrity attacks on integrated volt/VAR control and countermeasures," in *American Control Conference (ACC)*, June 2014, pp. 4372–4378.

[27] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.