On Kalman Filtering with Compromised Sensors: Attack Stealthiness and Performance Bounds

Cheng-Zong Bai, Vijay Gupta, and Fabio Pasqualetti

***Abstract*—Control systems operate under the assumption that sensors are trustworthy. Yet, when communication channels are unprotected or sensors are accessible from networked stations, malicious users can compromise the system by spoofing the measured information. We consider a linear time-invariant system with a single sensor, where the state is estimated by a Kalman filter. We assume the presence of an attacker with the ability to modify the measurements arbitrarily, which are then processed by the Kalman filter for as long as the attacker remains undetected. The objective of the attacker is to maximize the mean square error of the Kalman filter. We adopt a notion of attack stealthiness based on the Kullback–Leibler divergence measure, and characterize the worst case degradation induced by an attacker with a fixed stealthiness level. Additionally, we characterize optimal attack strategies that achieve our bound of performance degradation, thereby proving tightness of our result.**

***Index Terms*—Detection of stealthy malicious attacks, Kullback-Leibler divergence, security of cyberphysical systems.**

I. Introduction

Cyber-physical systems are an integral part of modern society, and need to operate reliably in the face of accidental and malicious malfunctions. Existing protection mechanisms based on data encryption and fault detection have proven ineffective especially against deliberate manipulation from resourceful attackers [2]–[4], showing the need for new theoretical and practical approaches to cyber-physical security [5].

For a system to function reliably, contingencies and malfunctions need to be promptly detected and remediated. While accidental malfunctions can be detected more easily [6], [7], malicious attacks can remain unnoticed when the system parameters and measurements are properly manipulated [8], [9], thereby posing additional challenges and risks. For deterministic systems, recent studies have shown that attack detectability is equivalent to the control-theoretic notion of *invariant zeros* [10]. In particular, in a deterministic system an attack is undetectable from the system measurements if and only if it excites only the zero dynamics of the attack-output system [8], [11]. Thus, attack detectability in deterministic systems has a binary answer that is independent of the detection algorithm. For systems driven by process and measurement noise, instead, attack detectability is typically defined based on common detection algorithms, such as the bad data detector [12]. While convenient for analysis, this approach fails to provide a comprehensive characterization of attack detectability. In fact, when the attacker has knowledge of the detection algorithm, it may orchestrate attacks that are undetectable from the algorithm that is being used, but could possibly be revealed with other methods. This motivates us to adopt a notion of attack stealthiness that does not rely on any specific type of detector, and that consequently allows us to reveal fundamental detectability properties.

In [1], we introduced the notion of ϵ-marginal stealthiness to quantify the stealthiness level in an estimation problem with respect to the class of ergodic detectors. This notion, however, is not sufficiently general and it lacks a concrete connection with useful detection metrics such as the error probability. Following [13], in this paper we adopt a notion of ϵ-stealthiness that is based on the information-theoretic notion of the Kullback–Leibler divergence (KLD), which quantifies the achievable exponent of the probability of false alarms and is independent of the attack-detection algorithm being used. We consider linear time-invariant systems with a single sensor, where the state is estimated by a Kalman filter. We allow the attacker to arbitrarily manipulate the measurements, with the objective to maintain a desired stealthiness level while maximizing the mean square error (MSE) of the Kalman filter implemented by the estimator with the corrupted measurements. Our analysis and results depart from the literature in different ways. For instance, compared to [14], we do not assume that the system is k-sparse observable, that is, we allow all sensors to be compromised by the attacker. Compared to [15], we do not restrict the detection scheme to the class of χ^2 detectors (in fact, we do not restrict the detection scheme to any particular class), and we do not restrict the attack strategies to be linear. Finally, we remark that since the submission of this paper, some recent literature has appeared that builds on it and uses a notion of attack detectability that is similar to what we propose in [1], [13], and in this paper. For instance, Kung *et al.* [16] extend the notion of ϵ-stealthiness given in [13] to higher order systems, and show how the performance of the attacker differ in the scalar and vector cases. In [17], Zhang and Venkitasubramaniam extend the setup in [13] to vector and not necessarily stationary systems, but consider a finite horizon problem. Two other relevant recent works are [18] that uses the notion of the KLD as a causal measure of information flow to quantify the effect of attacks on the system output, whereas [19] characterizes optimal attack strategies with respect to a linear quadratic cost that combines attackers control and undetectability goals.

The main contributions of this paper are twofold. First, we propose an information-theoretic approach to define a graded notion of

Manuscript received August 8, 2016; revised August 9, 2016 and April 3, 2017; accepted June 8, 2017. Date of publication June 13, 2017; date of current version December 1, 2017. The work of C.-Z. Bai was supported by the NSF Award CNS-1239224 and AFOSR Award FA9550-15-1-0186. The work of V. Gupta was supported by the NSF Awards CNS-1544724 and ECCS-1550016. The work of F. Pasqualetti was supported by the ONR Award #N00014-14-1-0816. Recommended by Associate Editor M. Verhaegen. This paper was presented in part at the IEEE American Control Conference, Portland, OR, USA, June 2014. *(Corresponding author: Vijay Gupta.)*

C.-Z. Bai and V. Gupta are with the Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: cbai@alumni.nd.edu; vgupta2@nd.edu).

F. Pasqualetti is with the Department of Mechanical Engineering, University of California at Riverside, Riverside, CA 92521 USA (e-mail: fabiopas@engr.ucr.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2017.2714903

attack stealthiness, namely ϵ-stealthiness, and to characterize fundamental limitations for the detection of sensor attacks in stochastic control systems. We derive limitations on the degradation of the MSE of the Kalman filter for strictly stealthy attacks ($\epsilon = 0$) and for general ϵ-stealthy attacks ($\epsilon > 0$). Furthermore, we quantify performance degradation as a function of the attacker's knowledge of the system, and as a tradeoff with the stealthiness level. Second, we design optimal attacks that achieve the identified performance bounds, and prove their tightness. Finally, we illustrate our results.

Notation: We denote a random variable by boldface $\mathbf{X}$ and its realization by the normal font x. The probability density function of $\mathbf{X}$ is denoted by $f_{\mathbf{X}}(x)$ and, by abusing the notation, by $f_{\mathbf{X}}$ or f_x. A Gaussian distribution with mean μ and variance σ^2 is denoted by $\mathcal{N}(\mu, \sigma^2)$. A sequence $\{y_n\}_{n=1}^k$ is denoted by y_1^k. The KLD between two random sequences x_1^k and y_1^k is defined as

$$D\big(x_1^k \big\| y_1^k\big) = \int_{-\infty}^{\infty} f_{x_1^k}(t_1^k) \log \frac{f_{x_1^k}(t_1^k)}{f_{y_1^k}(t_1^k)} dt_1^k \tag{1}$$

where $f_{x_1^k}$ and $f_{y_1^k}$ are the probability density functions of x_1^k and y_1^k, respectively. The set of real numbers is $\mathbb{R}$. $\operatorname{sgn}(\cdot)$ is the sign function with $\operatorname{sgn}(x) = 1$ if $x \geq 0$ and $\operatorname{sgn}(x) = -1$ if $x < 0$.

II. Problem Formulation

A. System Model

Consider a process with state $x_k \in \mathbb{R}$ that evolves as

$$\begin{aligned} x_{k+1} &= a x_k + w_k, \qquad k \geq 1, \\ y_k &= c x_k + v_k \end{aligned} \tag{2}$$

where $a, c \in \mathbb{R}$, $|a| < 1$, the initial condition $x_1 \sim \mathcal{N}(0, \Pi_0)$, and w_1^∞ and v_1^∞ represent the process noise sequence and the measurement noise sequence, respectively. Both noise sequences are assumed to be white processes with $w_k \sim \mathcal{N}(0, \sigma_w^2)$ and $v_k \sim \mathcal{N}(0, \sigma_v^2)$, with σ_w^2 and σ_v^2 being positive. All the random variables in the process noise sequence, measurement noise sequence, and the initial condition are assumed to be mutually independent.

If no attacker is present, an estimator uses the measurements y_1^k to generate a minimum MSE (MMSE) estimate $\hat{x}_{k+1}$ of the state x_{k+1} based on these measurements. The Kalman filter provides a recursive calculation for the estimate that minimizes the MSE $\mathbb{E}[(\hat{x}_{k+1} - x_{k+1})^2]$. Thus, the estimate evolves as

$$\hat{x}_{k+1} = a\hat{x}_k + K(k)(y_k - c\hat{x}_k) \tag{3}$$

where $K(k)$ is the Kalman gain. For nonzero a and c, the Kalman gain converges exponentially. In the sequel, we assume that the MMSE estimate is obtained by a steady-state Kalman filter with initial condition $\hat{x}_1 = 0$. The results in this paper can be generalized to more general initial conditions at the expense of more notation; the main intuition is that the problem formulation and the main results below are dominated by the steady state of the system. Given this assumption, we remove the time dependence on the Kalman gain and denote it by K. The Kalman gain is given by $K = acP(c^2P + \sigma_v^2)^{-1}$, where $P = \mathbb{E}[(\hat{x}_{k+1} - x_{k+1})^2]$ is the MSE of the state estimation, which is the positive solution to the following equation:

$$P = a^2 P + \sigma_w^2 - \frac{a^2 c^2 P^2}{c^2 P + \sigma_v^2}.$$

Let $z_k = y_k - c\hat{x}_k$ be the innovation of the Kalman filter at time k. Notice that z_1^∞ is a white sequence with $z_k \sim \mathcal{N}(0, \sigma_z^2)$ and $\sigma_z^2 = c^2 P + \sigma_v^2$.

B. Attack Model

An attacker can possibly replace the measurement sequence y_1^∞ transmitted by the sensor with any arbitrary attack sequence $\tilde{y}_1^\infty$. If the estimator is not aware of the presence of the attacker, the attack sequence $\tilde{y}_1^\infty$ is treated as the input to the Kalman filter. Denote the corresponding output of the Kalman filter by $\hat{\tilde{x}}_1^\infty$. This sequence is treated as the estimate of the state since the estimator does not know that an attack is in progress. Similar to (3), the sequence $\hat{\tilde{x}}_1^\infty$ is obtained as

$$\hat{\tilde{x}}_{k+1} = a\hat{\tilde{x}}_k + K(\tilde{y}_k - c\hat{\tilde{x}}_k) \tag{4}$$

where the initial condition $\hat{\tilde{x}}_1 = \hat{x}_1$. With this corrupted estimate, the estimation error is given by $(\hat{\tilde{x}}_{k+1} - x_{k+1})$. Denote the corresponding MSE that is induced by the attacker by $\tilde{P}_{k+1} = \mathbb{E}[(\hat{\tilde{x}}_{k+1} - x_{k+1})^2]$.

We now specify the information available at the attacker for the design the attack sequence $\tilde{y}_1^\infty$. We assume that the attacker knows the system model in (2), and that the information about the system variables at time k is denoted by the set $\mathcal{I}_k$. Examples of information patterns $\mathcal{I}_k$ are as follows.

1) The attacker has access to the system state ($\mathcal{I}_k = \{x_1^k\}$).
2) The attacker has access to the measurements ($\mathcal{I}_k = \{y_1^k\}$).
3) The attacker has access to the measurements with a delay $d \in \mathbb{N}$ ($\mathcal{I}_k = \{y_1^{k-d}\}$).
4) The attacker has no information about the state $\mathcal{I}_k = \{\emptyset\}$.

The attacker uses a Kalman filter to obtain an MMSE state estimate $\hat{x}_{k+1}^A$ of the state x_{k+1}. Let K_A be the steady-state Kalman gain in this filter, P_A its steady-state MSE, and $\{z_n^A\}_{n=1}^k$ the innovation at the attacker.

Assumption 1: Due to causality constraints, $\mathcal{I}_k$ is independent of w_k^∞ and v_{k+1}^∞, and the innovation $\{z_n^A\}_{n=1}^k$ is a white Gaussian process with $z_k^A \sim \mathcal{N}(0, \sigma_{z^A}^2)$. ■

The next result follows from the principle of orthogonality [20].

Lemma 1: For any information pattern $\mathcal{I}_1^\infty$ that satisfies Assumption 1,

1) the attacker's innovation z_k^A is independent of all random variables z_h^A, with $h < k$, generated by $\mathcal{I}_1^{k-1}$; and
2) the attacker's estimation error $(\hat{x}_{k+1}^A - x_{k+1})$ is independent of all random variables generated by $\mathcal{I}_k$.

Finally, denote the "innovation" sequence at the estimator in the presence of an attacker by $\tilde{z}_k = \tilde{y}_k - c\hat{\tilde{x}}_k$. Note that the sequence $\tilde{z}_1^\infty$ need not be i.i.d. nor does the marginal distribution of any random variable $\tilde{z}_k$ be Gaussian with mean 0 or variance σ_z^2. Since there is a bijective mapping between $\tilde{y}_1^k$ and $\tilde{z}_1^k$ for all $k \in \mathbb{N}$, we may call equivalently the sequence $\tilde{z}_1^\infty$ as the attack. We will make the following assumption.

Assumption 2: The innovation sequence $\{z_k^A\}_{k=1}^\infty$ is a minimal sufficient statistic [21] for the attack sequence $\tilde{z}_1^\infty$. In particular, this implies that the sequence $\{z_i^A\}_{i=1}^{n-1}$ can be reconstructed from the sequence $\tilde{z}_1^{n-1}$. Since $\tilde{z}_1^{n-1}$ is a function of the attacker's information pattern $\{z_i^A\}_{i=1}^{n-1}$, the two sequences thus have a bijective mapping. ■

Assumption 2 implies that the information pattern at the attacker can equivalently be represented as $\mathcal{I}_k = \{z_n^A\}_{n=1}^k$.

C. Stealthiness

The attacker is constrained in the input $\tilde{y}_1^\infty$ it replaces since it seeks to be stealthy or undetected by the controller. If the estimator is aware

that an attacker has replaced the correct measurement sequence y_1^∞ by a different sequence $\tilde{y}_1^\infty$, the system can presumably switch to a safer mode of operation. Notions of stealthiness have been proposed in the literature. For deterministic closed-loop systems, Pasqualetti *et al.* [8] showed that stealthiness of an attacker is equivalent to the existence of zero dynamics for the system driven by the attack. Similar to [8], we seek a notion of stealthiness without placing any restrictions on the attacker or the detector employed by the estimator, but for stochastic systems in an estimation context.

To this end, we follow the development in [13] and pose the problem of detecting an attacker by the estimator as a (sequential) hypothesis testing problem. Specifically, the estimator relies on the received measurements to decide the following binary hypothesis testing problem:

$$H_0 : \text{ No attack is in progress (the estimator receives } y_1^k);$$

$$H_1 : \text{ Attack is in progress (the estimator receives } \tilde{y}_1^k).$$

For a given detector employed at the estimator to select one of the two hypotheses, denote the probability of false alarm (i.e., the probability of deciding H_1 when H_0 is true) at time k by p_k^F, and the probability of correct detection (i.e., the probability of deciding H_1 when H_1 is true) at time k by p_k^D.

One may envisage that stealthiness of an attacker implies $p_k^D = 0$. However, as is standard in detection theory, we need to consider both the quantities p_k^F and p_k^D simultaneously.[1] Intuitively, an attack is hard to detect if the performance of *any* detector is independent of the received measurements. Thus, we define an attacker to be stealthy if there exists no detector that can perform better (in the sense of simultaneously achieving higher p_k^D and lower p_k^F) than a detector that makes a decision by ignoring all the measurements and making a random guess to decide between the hypotheses.

Definition 1 (Stealthy attacks [13]): An attack $\tilde{y}_1^\infty$ is

1) strictly stealthy, if there exists no detector such that $p_k^F < p_k^D$ for any $k > 0$.
2) ϵ-stealthy, if, for $\epsilon > 0$ and any value of $\Delta \in (0, 1)$, there exists no detector for which $0 < 1 - p_k^D \leq \Delta$ and

$$\limsup_{k\to\infty} -\frac{1}{k} \log p_k^F > \epsilon. \tag{5}$$

■

Intuitively, an attack is strictly stealthy if no detector can perform better than a random guess in deciding whether an attack is in progress. Furthermore, an attack is ϵ-stealthy if there exists no detector such that $0 < 1 - p_k^D \leq \delta$ for all time k and p_k^F converges to zero exponentially fast with rate greater than ϵ as $k \to \infty$.

D. Performance Metric

We assume that the attacker aims at maximizing the MSE $\tilde{P}(k+1)$ for the estimate calculated at the estimator. To remove the dependence on a particular time k, we consider the asymptotic behavior of $\tilde{P}_{k+1}$. Specifically, the metric of the performance degradation that the attacker can induce is the limit superior to the time averaged MSE, as given by

$$\tilde{P} \triangleq \limsup_{k\to\infty} \frac{1}{k} \sum_{n=0}^{k} \tilde{P}_{n+1}. \tag{6}$$

This metric exists even for those attack sequences for which $\tilde{P}_{k+1}$ does not converge. Note that if $\tilde{P}_1^\infty$ is a convergent sequence, then $\tilde{P} = \lim_{k\to\infty} \tilde{P}_k$ (see, e.g., [21]).

[1]For instance, a detector that always declares H_1 to be true will achieve $p_k^D = 1$. However, it will not be a good detector because $p_k^F = 1$.

We seek to solve the following problems.

1) What is the performance degradation that a strictly stealthy attack can induce? What is such an attack?
2) What is the performance degradation that an ϵ-stealthy attack can induce? What is such an attack?

III. Main Results

A. Preliminary Results

The following results can be proved along the lines of [13] and relate stealthiness to the KLD, thus providing an operational definition that is easier to work with.

Lemma 2 (Condition for Strictly Stealthiness): An attack $\tilde{z}_1^\infty$ is strictly stealthy if and only if $\tilde{z}_1^\infty$ is a sequence of i.i.d. Gaussian random variables where $\tilde{z}_k \sim \mathcal{N}(0, \sigma_z^2)$. ■

Lemma 3 (Conditions for ϵ-stealthiness): If an attack $\tilde{z}_1^\infty$ is ϵ-stealthy, then the following condition holds:

$$\limsup_{k\to\infty} \frac{1}{k} D\big(\tilde{z}_1^k \big\| z_1^k\big) \leq \epsilon. \tag{7}$$

Conversely, if an attack sequence $\tilde{z}_1^\infty$ is ergodic and satisfies

$$\lim_{k\to\infty} \frac{1}{k} D\big(\tilde{z}_1^k \big\| z_1^k\big) \leq \epsilon \tag{8}$$

then the attack is ϵ-stealthy. ■

B. Strictly Stealthy Attacks

We begin by considering the performance degradation induced by a strictly stealthy attack. Define

$$\tilde{e}_{k+1} \triangleq \hat{\tilde{x}}_{k+1} - \hat{x}_{k+1}^A.$$

Lemma 4: The MSE $\tilde{P}(k+1)$ can be expressed as

$$\tilde{P}_{k+1} = P_A + \mathbb{E}\big[\tilde{e}_{k+1}^2\big]. \tag{9}$$

Proof: First, note that $\mathbb{E}[\tilde{e}_{k+1}(\hat{x}_{k+1}^A - x_{k+1})] = 0$ because of the orthogonality principle. Thus, we can write

$$\begin{aligned}\tilde{P}_{k+1} &= \mathbb{E}\big[\big(\hat{\tilde{x}}_{k+1} - \hat{x}_{k+1}^A + \hat{x}_{k+1}^A - x_{k+1}\big)^2\big] \\ &= P_A + \mathbb{E}\big[\tilde{e}_{k+1}^2\big] + 2\mathbb{E}\big[\tilde{e}_{k+1}(\hat{x}_{k+1}^A - x_{k+1})\big] \\ &= P_A + \mathbb{E}\big[\tilde{e}_{k+1}^2\big].\end{aligned}$$

■

Lemma 5: For a strictly stealthy attack $\tilde{z}_1^\infty$, $\mathbb{E}[\hat{\tilde{x}}_{k+1}^2] = \mathbb{E}[\hat{x}_{k+1}^2]$.

Proof: Linear recursions (3) and (4) that generate the estimates $\hat{x}_{k+1}$ and $\hat{\tilde{x}}_{k+1}$, respectively, are identical except for the driving terms being z_1^k (for $\hat{x}_{k+1}$) and $\tilde{z}_1^k$ (for $\hat{\tilde{x}}_{k+1}$). Now, note that if the attack $\tilde{z}_1^\infty$ is strictly stealthy, Lemma 2 implies that similar to z_1^k, $\tilde{z}_1^k$ is also an i.i.d. sequence of Gaussian random variables with mean zero and variance σ_z^2. Since the initial conditions for recursions (3) and (4) are also identical, we have $\mathbb{E}[\hat{\tilde{x}}_{k+1}^2] = \mathbb{E}[\hat{x}_{k+1}^2]$. ■

We now have the following result.

Theorem 1 (Performance degradation of strictly stealthy attacks): Consider the problem formulation in Section II. For any strictly stealthy attack $\tilde{z}_1^\infty$, the MSE $\tilde{P}$ induced satisfies

$$\tilde{P} \leq P_A + \big((\sigma_x^2 - P)^{\frac{1}{2}} + (\sigma_x^2 - P_A)^{\frac{1}{2}}\big)^2 \tag{10}$$

where $\sigma_x^2 \triangleq \lim_{k\to\infty} \mathbb{E}[x_k^2] = \frac{\sigma_w^2}{1-a^2}$. Moreover, one strictly stealthy attack $\tilde{z}_1^\infty$ that achieves the upper bound is given by

$$\tilde{z}_k = -\sqrt{\frac{\sigma_z^2}{\sigma_{z^A}^2}}\, z_k^A. \tag{11}$$

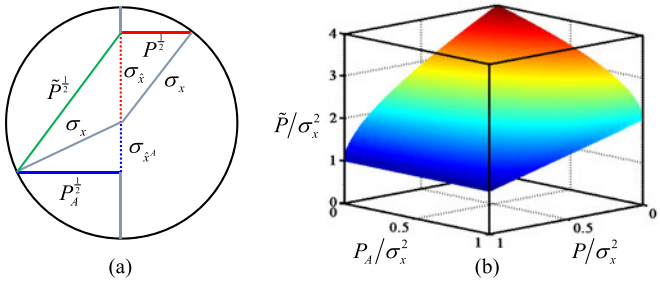

Fig. 1. (a) Geometric interpretation of Theorem 1, and (b) performance degradation $\tilde{P}$ induced by a strictly stealthy attack as a function of P and P_A.

Proof: We begin by noticing that

$$\begin{aligned}\mathbb{E}\big[\tilde{e}_{k+1}^2\big] &= \mathbb{E}\big[(\hat{\tilde{x}}_{k+1} - \hat{x}_{k+1}^A)^2\big] \\ &\overset{(a)}{\leq} \big(\mathbb{E}\big[\hat{\tilde{x}}_{k+1}^2\big]^{\frac{1}{2}} + \mathbb{E}\big[(\hat{x}_{k+1}^A)^2\big]^{\frac{1}{2}}\big)^2 \\ &\overset{(b)}{=} \big(\mathbb{E}\big[\hat{x}_{k+1}^2\big]^{\frac{1}{2}} + \mathbb{E}\big[(\hat{x}_{k+1}^A)^2\big]^{\frac{1}{2}}\big)^2 \\ &\overset{(c)}{=} \big(\big(\mathbb{E}\big[x_{k+1}^2\big] - P\big)^{\frac{1}{2}} + \big(\mathbb{E}\big[x_{k+1}^2\big] - P_A\big)^{\frac{1}{2}}\big)^2 \end{aligned} \tag{12}$$

where (a) follows from Minkowski's inequality [22, Th. 4.7.5], (b) follows from Lemma 5, and (c) follows from the principle of orthogonality since $\hat{x}_{k+1}$ and $\hat{x}_{k+1}^A$ are both MMSE estimates for x_{k+1}. The upper bound now follows by substituting (12) into (9) and considering the limit for $k \to \infty$.

For achievability, we need to identify the conditions for Minkowski's inequality above to hold with equality. For this, the attack must be such that $\hat{\tilde{x}}_{k+1} = -\beta \hat{x}_{k+1}^A$ for a given constant $\beta \geq 0$. This condition is satisfied by the attack (11) given the linearity of the Kalman filter recursions. Thus, we are left to verify if the attack is strictly stealthy. Since, with this attack $\tilde{z}_1^\infty$ is an i.i.d. Gaussian sequence with $\tilde{z}_k \sim \mathcal{N}(0, \sigma_z^2)$, this condition is also satisfied, and the theorem follows. ∎

Remark 1 (Geometric illustration of Theorem 1): The upper bound in Theorem 1 can be illustrated geometrically, as shown in Fig. 1. In fact, from (10) and (11) we have

$$\tilde{P}^{\frac{1}{2}} = \left(\left(P_A^{\frac{1}{2}}\right)^2 + (\sigma_{\hat{x}^A} + \sigma_{\hat{x}})^2\right)^{\frac{1}{2}}$$

where $\sigma_{\hat{x}^A} = (\sigma_x^2 - (P_A^{\frac{1}{2}})^2)^{\frac{1}{2}}$ and $\sigma_{\hat{x}} = (\sigma_x^2 - (P^{\frac{1}{2}})^2)^{\frac{1}{2}}$. We observe that, because σ_x is constant, if a strict stealthy attacker has more information about the state variable (i.e., with a smaller value of P_A), it can induce larger MSE $\tilde{P}$. Similarly, if the state estimator believes that the received data are trustworthy (i.e., with a small value of P), then a strictly stealthy attacker can induce larger error. ∎

C. ϵ-Stealthy Attacks

We now characterize the performance limitations of ϵ-stealthy attacks. We first present a converse result that gives an upper bound for the MSE $\tilde{P}$ induced by an ϵ-stealthy attack. Then, we provide an ϵ-stealthy attack that achieves this bound.

1) Preliminary Technical Results: We will use the following technical results in the later derivations. We refer the interested reader to the Appendix for a proof of these results. The first result follows immediately from the monotonicity and concavity of the function $f(x) = \sqrt{x}$.

Lemma 6: Suppose that $x \leq c_1 + c_2\sqrt{x}$, where $x, c_1,$ and c_2 are nonnegative real numbers. Then

$$x \leq \frac{2c_1 + c_2^2 + \sqrt{(2c_1 + c_2^2)^2 - 4c_1^2}}{2} \tag{13}$$

where the upper bound is the unique solution to the equation $x = c_1 + c_2\sqrt{x}$. Moreover, the upper bound is monotonically increasing with respect to c_1 and c_2.

The second result provides some bounds that will be used in the later results. Let $\alpha_n a\tilde{e}_n$ be the linear MMSE (LMMSE) estimate of $K\tilde{z}_n$ given $a\tilde{e}_n$ and let M_n be its corresponding MSE. Also, define the quantities

$$\begin{aligned}\mathcal{E}_k &\triangleq \frac{1}{k}\sum_{n=1}^k \mathbb{E}\big[\tilde{z}_n^2\big] \\ \beta_k &\triangleq \left(\frac{1}{\sigma_{z^A}^2 k}\sum_{n=1}^k M_n\right)^{\frac{1}{2}} \geq 0. \end{aligned} \tag{14}$$

Lemma 7: For any time n, the following inequalities hold:

$$\frac{1}{k}\sum_{n=1}^k \mathbb{E}\big[K\tilde{z}_n K_A z_n^A\big] \leq \sigma_{z^A}\beta_k\big(K_A^2\sigma_{z^A}^2\big)^{\frac{1}{2}} \tag{15}$$

$$\frac{1}{k}\sum_{n=1}^k \mathbb{E}\big[a\tilde{e}_n K\tilde{z}_n\big] \leq \left(K^2\mathcal{E}_k - (\sigma_{z^A}\beta_k)^2\right)^{\frac{1}{2}}\left(\frac{1}{k}\sum_{n=1}^k a^2\mathbb{E}\big[\tilde{e}_n^2\big]\right)^{\frac{1}{2}}. \tag{16}$$

Furthermore, (15) holds with equality if $\mathbb{E}[K\tilde{z}_n K_A z_n^A] > 0$, z_n^A is a scalar multiple of $K\tilde{z}_n - \alpha_n a\tilde{e}_n$, and the sequence $\tilde{z}_1^\infty$ is stationary. Similarly (16) holds with equality if $\mathbb{E}[a\tilde{e}_n K\tilde{z}_n] > 0$ and the sequence $\tilde{z}_1^\infty$ is stationary.

We can now bound two important quantities.

Lemma 8: The time average of $\mathbb{E}[\tilde{e}_{n+1}^2]$ is bounded as

$$\frac{1}{k}\sum_{n=1}^k \mathbb{E}\big[\tilde{e}_n^2\big] \leq \frac{S_k + \sqrt{S_k^2 - 4(1-a^2)^2R_k^2}}{2(1-a^2)^2} \tag{17}$$

where

$$\begin{aligned}R_k &= K^2\mathcal{E}_k + K_A^2\sigma_{z^A}^2 + 2|K_A|\sigma_{z^A}^2\beta_k + \frac{1}{k}\mathbb{E}\big[\tilde{e}_1^2\big], \\ S_k &= 2(1-a^2)R_k + 4a^2\big(K^2\mathcal{E}_k - \sigma_{z^A}^2\beta_k^2\big).\end{aligned}$$

Lemma 9: The following relations hold

$$M_n \geq \frac{1}{2\pi e}e^{2h(K\tilde{z}_n|a\tilde{e}_n)} \tag{18}$$

$$\frac{1}{k}\sum_{n=1}^k h(\tilde{z}_n|\tilde{z}_1^{n-1}) \leq \frac{1}{2}\log\frac{2\pi e\sigma_{z^A}^2\beta_k^2}{K^2}. \tag{19}$$

2) Converse: We will use the following in the converse result.

Lemma 10: For any $\gamma > 0$, the following functions always exist:

$$\begin{aligned}\underline{\delta}(\gamma) &= \arg\min_{x\in\mathbb{R}} x, \\ &\quad \text{subject to } \frac{1}{2}x - \gamma - \frac{1}{2} \leq \frac{1}{2}\log x \\ \bar{\delta}(\gamma) &= \arg\max_{x\in\mathbb{R}} x, \\ &\quad \text{subject to } \frac{1}{2}x - \gamma - \frac{1}{2} \leq \frac{1}{2}\log x.\end{aligned}$$

Furthermore, $\underline{\delta} : [0,\infty] \to (0,1]$ and $\bar{\delta} : [0,\infty] \to [1,\infty)$.

Proof: Since a logarithm function is concave, the feasible region of x in the constraint $\frac{1}{2}x-\gamma-\frac{1}{2}\leq\frac{1}{2}\log x$ is a closed interval lower bounded by $\underline{\delta}(\gamma)$ and upper bounded by $\bar{\delta}(\gamma)$, as defined above. Thus, the result follows. ∎

We now present a converse result for the MSE induced by ϵ-stealthy attacks.

Theorem 2 (Converse): Consider the problem formulation from Section II. The MSE induced by any ϵ-stealthy attack is upper bounded by

$$\tilde{P}\leq P_A+\max_{\beta\geq 0}\frac{S+\sqrt{S^2-4(1-a^2)^2R^2}}{2(1-a^2)^2}$$
$$\text{s.t. }\underline{\delta}(\epsilon)K^2\sigma_z^2\leq\sigma_{z^A}^2\beta^2\leq\bar{\delta}(\epsilon)K^2\sigma_z^2 \tag{20}$$

where

$$R=K^2\mathcal{E}+K_A^2\sigma_{z^A}^2+2|K_A|\sigma_{z^A}^2\beta,$$
$$S=2(1-a^2)R+4a^2\big(K^2\mathcal{E}-\sigma_{z^A}^2\beta^2\big),$$
$$\mathcal{E}=\Big(2\epsilon+\log\frac{\sigma_{z^A}^2\beta^2}{K^2\sigma_z^2}+1\Big)\sigma_z^2. \tag{21}$$

Proof: We first prove the theorem for a finite number of time steps k and then let $k\to\infty$. From Lemma 8, the MSE induced by any attack is upper bounded as

$$\frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\big[\tilde{e}_n^2\big]\leq\frac{S_k+\sqrt{S_k^2-4(1-a^2)^2R_k^2}}{2(1-a^2)^2} \tag{22}$$

where

$$R_k=K^2\mathcal{E}_k+K_A^2\sigma_{z^A}^2+2|K_A|\sigma_{z^A}^2\beta_k+\frac{1}{k}\mathbb{E}\big[\tilde{e}_1^2\big],$$
$$S_k=2(1-a^2)R_k+4a^2\big(K^2\mathcal{E}_k-\sigma_{z^A}^2\beta_k^2\big).$$

Thus, we seek to solve

$$\max_{\beta_k,\mathcal{E}_k}\frac{S_k+\sqrt{S_k^2-4(1-a^2)^2R_k^2}}{2(1-a^2)^2} \tag{23}$$
$$\text{subject to attack being }\epsilon-\text{stealthy}.$$

From Lemma 3, for an ϵ-stealthy attack, condition (7) must hold. Assume for a finite k that $\frac{1}{k}D(\tilde{z}_1^k\|z_1^k)\leq\epsilon$. Since the innovation z_1^∞ is a sequence of i.i.d. $\mathcal{N}(0,\sigma_z^2)$ random variables, we can write

$$\frac{1}{k}D\big(\tilde{z}_1^k\big\|z_1^k\big)=\frac{1}{2}\log(2\pi\sigma_z^2)+\frac{\mathcal{E}_k}{2\sigma_z^2}-\frac{1}{k}\sum_{n=1}^{k}h(\tilde{z}_n|\tilde{z}_1^{n-1}). \tag{24}$$

Using Lemma 9 yields

$$\frac{\mathcal{E}_k}{2\sigma_z^2}=\frac{1}{k}D\big(\tilde{z}_1^k\big\|z_1^k\big)+\frac{1}{k}\sum_{n=1}^{k}h\big(\tilde{z}_n\big|\tilde{z}_1^{n-1}\big)-\frac{1}{2}\log\big(2\pi\sigma_z^2\big)$$
$$\leq\epsilon+\frac{1}{2}\log\frac{\sigma_{z^A}^2\beta_k^2}{K^2\sigma_z^2}+\frac{1}{2}. \tag{25}$$

We now translate this into constraints on β_k and $\mathcal{E}_k$ as follows.

1) Note that M_n is the MSE of the LMMSE estimate of $K\tilde{z}_n$, and thus, must satisfy $M_n\leq K^2\mathbb{E}[\tilde{z}_n^2]$. Using this relation with (14), we obtain

$$\sigma_{z^A}^2\beta_k^2=\frac{1}{k}\sum_{n=1}^{k}M_n\leq K^2\mathcal{E}_k.$$

Plugging this inequality in the inequality constraint (25) thus yields

$$\frac{\sigma_{z^A}^2\beta_k^2}{2K^2\sigma_z^2}\leq\epsilon+\frac{1}{2}\log\frac{\sigma_{z^A}^2\beta_k^2}{K^2\sigma_z^2}+\frac{1}{2}. \tag{26}$$

Using Lemma 10 thus implies that we can restrict the search of β_k to the region

$$\underline{\delta}(\epsilon)K^2\sigma_z^2\leq\sigma_{z^A}^2\beta_k^2\leq\bar{\delta}(\epsilon)K^2\sigma_z^2. \tag{27}$$

2) From Lemma 6, the objective function in (23) is monotonic increasing with respect to $\mathcal{E}_k$. Therefore, in order to maximize this function, inequality (25) must hold with equality. Thus, the choice of $\mathcal{E}_k$ is from the equality constraint

$$\mathcal{E}_k=\Big(2\epsilon+\log\frac{\sigma_{z^A}^2\beta_k^2}{K^2\sigma_z^2}+1\Big)\sigma_z^2. \tag{28}$$

Thus, (23) is equivalent to solving

$$\max_{\beta_k>0}\frac{S_k+\sqrt{S_k^2-4(1-a^2)^2R_k^2}}{2(1-a^2)^2} \tag{29}$$
$$\text{subject to (27) and (28)}.$$

Now, note that all the manipulations in the proof are continuous. Thus, we can let $k\to\infty$ and the corresponding limits of all quantities [in particular, the limit of the right-hand side in (22)] will exist. The proof now follows from (9) and the fact that $\frac{1}{k}\mathbb{E}[\tilde{e}_1^2]\to 0$ as $k\to\infty$. ∎

3) Achievability: We now show that the upper bound presented in Theorem 2 can be achieved by an ϵ-stealthy attack. For notational simplicity, we denote the upper bound for $\tilde{P}$ in (22) by $\tilde{P}_{\max}$ and the maximizing value of β in (22) by $\beta^\star$.

Theorem 3 (Achievability): For any given ϵ, consider the attack $\tilde{z}_1^\infty$ generated using the following recursion:

$$\tilde{z}_k=a(1+\alpha^\star)\tilde{z}_{k-1}+\frac{a(\beta^\star\operatorname{sgn}(K_A)-\alpha^\star K_A)}{K}z_{k-1}^A$$
$$-\frac{\beta^\star\operatorname{sgn}(K_A)}{K}z_k^A \tag{30}$$

with the initial conditions $\tilde{z}_0=\tilde{z}_0^A=0$

$$\alpha^\star\triangleq\sqrt{\frac{K^2\mathcal{E}^\star-\sigma_{z^A}^2(\beta^\star)^2}{a^2(\tilde{P}_{\max}-P_A)}}$$
$$\mathcal{E}^\star\triangleq\Big(2\epsilon+\log\frac{\sigma_{z^A}^2(\beta^\star)^2}{K^2\sigma_z^2}+1\Big)\sigma_z^2.$$

The attack $\tilde{z}_1^\infty$ induces MSE at the estimator equal to the upper bound $\tilde{P}_{\max}$ in Theorem 2 and is ϵ-stealthy.

Proof: The proof is described in three steps. First, we write the attack sequence in (30) in an alternate form that makes it easier to reason. Then, we show that the attack induces MSE equal to $\tilde{P}_{\max}$. Finally, we show that the attack is ϵ-stealthy.

To this end, we first prove that the attack $\tilde{z}_1^\infty$ can be considered to be generated using the following relation:

$$K\tilde{z}_k=\alpha^\star a\tilde{e}_k-\beta^\star\operatorname{sgn}(K_A)z_k^A. \tag{31}$$

This is because (31) directly yields

$$\begin{aligned} K\tilde{z}_k &= \alpha^\star a\Big(a(\hat{\tilde{x}}_{k-1} + K\tilde{z}_{k-1}) - (a\hat{x}^A_{k-1} + K_A z^A_{k-1})\Big) \\ &\quad - \beta^\star \operatorname{sgn}(K^A) z^A_k \\ &= a\big(\alpha^\star a\tilde{e}_{k-1} - \beta^\star \operatorname{sgn}(K_A) z^A_{k-1}\big) + \alpha^\star aK\tilde{z}_{k-1} \\ &\quad + \big(a\beta^\star \operatorname{sgn}(K_A) - \alpha^\star aK_A\big) z^A_{k-1} - \beta^\star \operatorname{sgn}(K^A) z^A_k \\ &= aK(1+\alpha^\star)\tilde{z}_{k-1} + a(\beta^\star \operatorname{sgn}(K_A) - \alpha^\star K_A) z^A_{k-1} \\ &\quad - \beta^\star \operatorname{sgn}(K^A) z^A_k, \\ \Rightarrow \tilde{z}_k &= a(1+\alpha^\star)\tilde{z}_{k-1} + \frac{a(\beta^\star \operatorname{sgn}(K_A) - \alpha^\star K_A)}{K} z^A_{k-1} \\ &\quad - \frac{\beta^\star \operatorname{sgn}(K_A)}{K} z^A_k \end{aligned}$$

which is identical to (30).

Now, we prove that the attack $\tilde{z}_1^\infty$ induces MSE equal to $\tilde{P}_{\max}$. First note from (31) that $\alpha_n = \alpha^\star$ and $\beta_n = \beta^\star \ \forall n \in \mathbb{N}$ for this attack. Since $-2\mathbb{E}[K\tilde{z}_n K_A z_n^A] = 2|K_A|\beta^\star \sigma^2_{z^A} \geq 0, 2\mathbb{E}[a\tilde{e}_n K\tilde{z}_n] = 2\alpha^\star a^2 \mathbb{E}[\tilde{e}_n^2] \geq 0$, z_n^A is a scalar multiplication of $K\tilde{z}_n - \alpha_n a\tilde{e}_n$, and the random sequence $\tilde{z}_1^\infty$ is stationary, from Lemma 7, both the relations (15) and (16) hold with equality. Consequently, Lemma 8 implies that (17) holds with equality. Finally, from the structure of (31), following the proof of Lemma 9, the relation (19) holds with equality. Thus, following the proof of Theorem 2, we obtain that the attack $\tilde{z}_1^\infty$ induces MSE equal to $\tilde{P}_{\max}$.

Finally, we can show that the attack $\tilde{z}_1^\infty$ generated by (30) is ϵ-stealthy as follows. Sequence (30) is a Gaussian autoregressive-moving-average (ARMA) sequence and hence its entropy rate is given by [23]

$$\lim_{k\to\infty} \frac{1}{k} \sum_{n=1}^{k} h\big(\tilde{z}_n \big| \tilde{z}_1^{n-1}\big) = \frac{1}{2} \log \frac{2\pi e(\beta^\star)^2 \sigma^2_{z^A}}{K^2}. \tag{32}$$

Furthermore, by definition

$$\begin{aligned} \lim_{k\to\infty} \mathcal{E}_k &= \lim_{k\to\infty} \mathbb{E}[\tilde{z}_k^2] \\ &= \frac{(\alpha^\star a)^2 (\tilde{P}_{\max} - P_A) + (\beta^\star)^2 \sigma^2_{z^A}}{K^2} = \mathcal{E}^\star. \end{aligned} \tag{33}$$

Using (24), (32), and (33), we thus obtain $\lim_{k\to\infty} \frac{1}{k} D(\tilde{z}_1^k \| z_1^k) = \epsilon$. Finally, a Gaussian ARMA sequence is ergodic. Thus, both the conditions stated in Lemma 3 are satisfied and the attack (31) is ϵ-stealthy. ■

4) Discussion: Some observations are in order. First, Theorems 2 and 3 characterize the minimum-mean-square estimation error achievable by an ϵ-stealthy attack as a function of the system parameters, noise statistics, and information available to the attacker. The two theorems provide a fundamental limitation for the estimation error induced by any ϵ-stealthy attack, in the sense that the bounds are independent of any specific detection mechanism employed by the estimator. Second, the maximum MSE $\tilde{P}_{\max}$ that can be induced by an ϵ-stealthy attacker is monotonically increasing with ϵ [see the upper bound in (20)]. Thus, an attacker that is less stealthy can induce a higher MSE. Third, it can be verified that the derivative of $\tilde{P}_{\max}$ with respect to P_A is negative. Thus, $\tilde{P}_{\max}$ decreases monotonically with respect to P_A, which is intuitively satisfying since it implies that if the attacker has more information about the state, then it can induce a larger MSE at the estimator. Finally, the optimal ϵ-stealthy attack $\tilde{z}_1^\infty$ is an ARMA Gaussian sequence. Thus, the random sequence $\tilde{z}_1^\infty$ is not white and using a residual error detector [4] may not be optimal to detect this attack in the sense of maximizing the exponent of the probability of false alarm or under the Neyman–Pearson criterion. The design of an optimal detector for this type of attacks is left as a direction of future research.

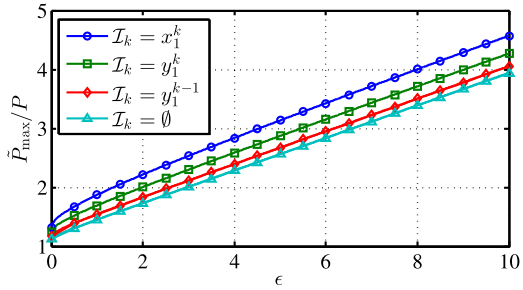

Fig. 2. Upper bound of $\tilde{P}$ in Theorem 2 as a function of ϵ for different information patterns at the attacker.

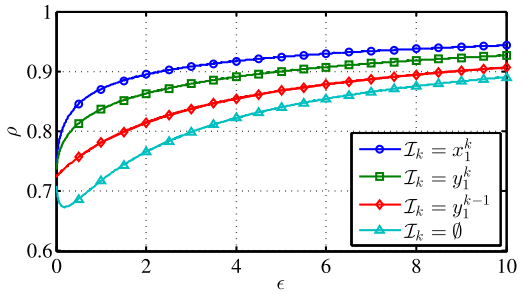

Fig. 3. Quantity ρ of the optimal attack (30) versus ϵ for different information patterns at the attacker.

IV. Numerical Results

To illustrate the results, consider a system as in (2) with parameters $a = 0.4$, $c = 1$, $\sigma_w^2 = 0.2$, and $\sigma_v^2 = 0.5$.

Induced MSE Versus the Level of Stealthiness: We first illustrate the tradeoff between an attacker's level of stealthiness and the induced MSE, as characterized in Theorem 2. Fig. 2 shows the upper bound $\tilde{P}_{\max}$ as a function of ϵ for different information patterns at the attacker. It can be seen that for any information pattern, the MSE induced by the attacker increases as its stealthiness decreases and it is more easily detected.

Memory of the Optimal Attack: Since z_1^∞ is an independent random sequence, we can write

$$\frac{1}{k} D\big(\tilde{z}_1^k \big\| z_1^k\big) = \frac{1}{k} \sum_{n=1}^{k} I\big(\tilde{z}_n; \tilde{z}_1^{n-1}\big) + \frac{1}{k} \sum_{n=1}^{k} D\big(\tilde{z}_n \| z_n\big)$$

where $I(\tilde{z}_n; \tilde{z}_1^{n-1})$ denotes the mutual information [21] between $\tilde{z}_n$ and $\tilde{z}_1^{n-1}$. An interesting interpretation of this equation is that the stealthiness of the attack as measured by $\frac{1}{k} D(\tilde{z}_1^k \| z_1^k)$ consists of two terms. The term $\frac{1}{k} \sum_{n=1}^{k} I(\tilde{z}_n; \tilde{z}_1^{n-1})$ characterizes the memory of the random sequence $\tilde{z}_1^\infty$, whereas the term $\frac{1}{k} \sum_{n=1}^{k} D(\tilde{z}_n \| z_n)$ measures the stealthiness obtained through the marginal pdf of $\tilde{z}_1^k$. For a stationary attack $\tilde{z}_1^\infty$, we define $\rho = \frac{\lim_{k\to\infty} D(\tilde{z}_k \| z_k)}{\lim_{k\to\infty} \frac{1}{k} D(\tilde{z}_1^k \| z_1^k)}$. The quantity ρ can be used to quantify the portion of stealthiness that is lost through the marginal pdf of $\tilde{z}_1^\infty$. If $\rho = 1$, then the attack $\tilde{z}_1^\infty$ is a memoryless random sequence. In particular, for the optimal attack, as characterized in [13], we have $\rho = 1$. If $\rho = 0$, then the attack $\tilde{z}_1^\infty$ is a colored Gaussian sequence with $\tilde{z}_k \sim \mathcal{N}(0, \sigma_z^2)$. Fig. 3 shows the quantity ρ for the optimal attack (30) versus the level of stealthiness ϵ for various

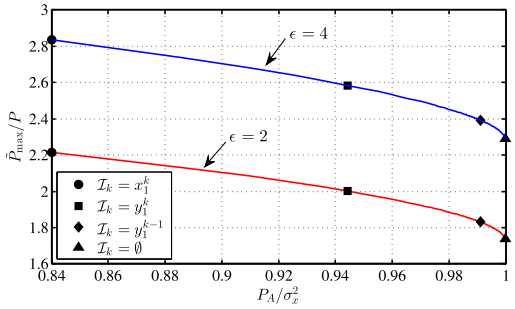

Fig. 4. Upper bound in Theorem 2 as a function of the attacker's information about the state.

information patterns at the attacker. As expected, since the optimal attack $\tilde{z}_1^\infty$ in (30) is an ARMA sequence, it has $\rho < 1$. Another interesting observation is that if the attacker can estimate the state x_k better, the optimal attack loses a higher fraction of stealthiness through the marginal pdf of $\tilde{z}_1^\infty$.

Induced MSE Versus the Attacker's Information Pattern: For a fixed level of stealthiness ϵ, Fig. 4 shows the MSE $\tilde{P}_{\max}$ for various values of P_A. A smaller value of P_A corresponds to better information about the state at the attacker, in which case it can induce higher error at the estimator.

V. Conclusion

We study a problem in which an attacker can compromise the measurements about the state of a linear time-invariant scalar process, which are transmitted from a sensor to the estimator. An information-theoretic notion of stealthiness that is independent of any specific detection mechanism employed by the estimator is considered. We analytically characterize the tradeoff between the stealthiness of the attacker and the maximal MSE for the state estimation at the estimator that the attack can induce. Moreover, we present an optimal attack that induces the maximal MSE. In particular, we show that an optimal ϵ-stealthy attack $\tilde{z}_1^\infty$ is a Gaussian ARMA random sequence.

Appendix

Proof of Lemma 7: To prove (15), note that

$$\frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\big[K\tilde{z}_n K_A z_n^A\big] \le \frac{1}{k}\sum_{n=1}^{k}\Big|\mathbb{E}\big[K\tilde{z}_n K_A z_n^A\big]\Big| \tag{A-1}$$

$$\overset{(a)}{=} \frac{1}{k}\sum_{n=1}^{k}\Big|\mathbb{E}\big[(K\tilde{z}_n - \alpha_n a\tilde{e}_n)K_A z_n^A\big]\Big|$$

$$\overset{(b)}{\le} \frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\big[(K\tilde{z}_n - \alpha_n a\tilde{e}_n)^2\big]^{\frac{1}{2}}\big(K_A^2\sigma_{z^A}^2\big)^{\frac{1}{2}} \tag{A-2}$$

$$\overset{(c)}{\le} \left(\frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\big[(K\tilde{z}_n - \alpha_n a\tilde{e}_n)^2\big]\right)^{\frac{1}{2}}\big(K_A^2\sigma_{z^A}^2\big)^{\frac{1}{2}}$$

$$\overset{(d)}{=} \left(\frac{1}{k}\sum_{n=1}^{k}M_n\right)^{\frac{1}{2}}\big(K_A^2\sigma_{z^A}^2\big)^{\frac{1}{2}} \overset{(e)}{=} \sigma_{z^A}\beta_k\big(K_A^2\sigma_{z^A}^2\big)^{\frac{1}{2}} \tag{A-3}$$

where (a) follows since both $\hat{\tilde{x}}_n$ and $\hat{x}_n^A$ (and hence also $\tilde{e}_n$) are linear functions of $\mathcal{I}_{n-1}$ and are thus independent of the attacker's innovation z_n^A at time n, (b) and (c) follow from application of the Cauchy–Schwarz inequality, (d) follows from the definition of M_n, and (e) follows from the definition of β_k. The adequacy of the stated conditions for (15) to hold with equality follows from the fact that they are sufficient for the inequalities mentioned above to hold with equality.

Then, we can write

$$\frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\big[a\tilde{e}_n K\tilde{z}_n\big] \le \frac{1}{k}\sum_{n=1}^{k}\Big|\mathbb{E}\big[a\tilde{e}_n K\tilde{z}_n\big]\Big|$$

$$\overset{(a)}{=} \frac{1}{k}\sum_{n=1}^{k}|\alpha_n|a^2\mathbb{E}\big[\tilde{e}_n^2\big] \tag{A.4}$$

$$\overset{(b)}{\le} \left(\frac{1}{k}\sum_{n=1}^{k}\alpha_n^2 a^2\mathbb{E}\big[\tilde{e}_n^2\big]\right)^{\frac{1}{2}}\left(\frac{1}{k}\sum_{n=1}^{k}a^2\mathbb{E}\big[\tilde{e}_n^2\big]\right)^{\frac{1}{2}}$$

$$\overset{(c)}{=} \left(K^2\mathcal{E}_k - \frac{1}{k}\sum_{n=1}^{k}M_n\right)^{\frac{1}{2}}\left(\frac{1}{k}\sum_{n=1}^{k}a^2\mathbb{E}\big[\tilde{e}_n^2\big]\right)^{\frac{1}{2}}$$

$$\overset{(d)}{=} \left(K^2\mathcal{E}_k - (\sigma_{z^A}\beta_k)^2\right)^{\frac{1}{2}}\left(\frac{1}{k}\sum_{n=1}^{k}a^2\mathbb{E}\big[\tilde{e}_n^2\big]\right)^{\frac{1}{2}} \tag{A.5}$$

where (a) follows from the principle of orthogonality given that $\alpha_n a\tilde{e}_n$ is the linear MMSE estimate of $K\tilde{z}_n$ given $a\tilde{e}_n$, (b) follows from the Cauchy–Schwarz inequality, (c) is true because the principle of orthogonality implies the relation $M_n = K^2\mathbb{E}[\tilde{z}_n^2] - \alpha_n^2 a^2\mathbb{E}[\tilde{e}_n^2]$ and (d) follows from the definition of β_k. Furthermore, if $\mathbb{E}[a\tilde{e}_n K\tilde{z}_n] > 0$ and the sequence $\tilde{z}_1^\infty$ is stationary, the inequalities in the above chain hold with equality. ■

Proof of Lemma 8: First, note that z_n^A is independent of $\tilde{e}_n$ and hence $\mathbb{E}[a\tilde{e}_n K_A z_n^A] = 0$. Thus,

$$\begin{aligned}\mathbb{E}\big[\tilde{e}_{n+1}^2\big] &= \mathbb{E}\big[\big(a\tilde{e}_n + K\tilde{z}_n - K_A z_n^A\big)^2\big]\\ &= a^2\mathbb{E}\big[\tilde{e}_n^2\big] + K^2\mathbb{E}\big[\tilde{z}_n^2\big] + K_A^2\sigma_{z^A}^2\\ &\quad - 2\mathbb{E}\big[K\tilde{z}_n K_A z_n^A\big] + 2\mathbb{E}\big[a\tilde{e}_n K\tilde{z}_n\big].\end{aligned}$$

Taking the time average of both sides yields

$$\begin{aligned}\frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\big[\tilde{e}_{n+1}^2\big] &= \frac{a^2}{k}\sum_{n=1}^{k}\mathbb{E}\big[\tilde{e}_n^2\big] + K^2\mathcal{E}_k + K_A^2\sigma_{z^A}^2\\ &\quad - \frac{2}{k}\sum_{n=1}^{k}\mathbb{E}\big[K\tilde{z}_n K_A z_n^A\big] + \frac{2}{k}\sum_{n=1}^{k}\mathbb{E}\big[a\tilde{e}_n K\tilde{z}_n\big].\end{aligned}$$

Thus, we can use (15) and (16) to obtain

$$\begin{aligned}(1-a^2)\left(\frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\big[\tilde{e}_n^2\big]\right) &\le K^2\mathcal{E}_k + K_A^2\sigma_{z^A}^2 + 2|K_A|\sigma_{z^A}^2\beta_k\\ &\quad + 2|a|\Big(K^2\mathcal{E}_k - \sigma_{z^A}^2\beta_k^2\Big)^{\frac{1}{2}}\left(\frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\big[\tilde{e}_n^2\big]\right)^{\frac{1}{2}} + \frac{1}{k}\mathbb{E}\big[\tilde{e}_1^2\big].\end{aligned}$$

Using Lemma 6, the proof is complete. ■

Proof of Lemma 9: Relation (18) follows from the maximum entropy theorem [21, Corollary 8.6.6] since M_n is the MSE of estimating $K\tilde{z}_n$ from $a\tilde{e}_n$. Now, note that $\{\hat{\tilde{x}}_n, \hat{x}_n^A\} \to \{z_i^A\}_{i=1}^{n-1} \to \tilde{z}_n$ is a Markov chain since $\hat{\tilde{x}}_n$, $\hat{x}_n^A$, and $\tilde{z}_n$ are all generated based on the attacker's information pattern $\mathcal{I}_{n-1}$, which is given by the attacker's

innovation sequence $\{z_i^A\}_{i=1}^{n-1}$, as stated in Assumption 1. The proof of (19) now follows from the following set of inequalities:

$$\frac{1}{2}\log\frac{2\pi e\sigma^2_{z^A}\beta_k^2}{K^2} \overset{(a)}{=} \frac{1}{2}\log\left(\frac{2\pi e}{K^2}\cdot\frac{1}{k}\sum_{n=1}^{k}M_n\right)$$

$$\overset{(b)}{\geq} \frac{1}{2}\log\left(\frac{2\pi e}{K^2}\cdot\frac{1}{k}\sum_{n=1}^{k}\frac{1}{2\pi e}e^{2h(K\tilde{z}_n|a\tilde{e}_n)}\right)$$

$$\overset{(c)}{=} \frac{1}{2}\log\left(\frac{1}{k}\sum_{n=1}^{k}e^{2h(\tilde{z}_n|a\tilde{e}_n)}\right) \overset{(d)}{\geq} \frac{1}{2}\log\left(\prod_{n=1}^{k}e^{2h(\tilde{z}_n|a\tilde{e}_n)}\right)^{\frac{1}{k}}$$

$$\overset{(e)}{\geq} \frac{1}{k}\sum_{n=1}^{k}h\big(\tilde{z}_n\big|\{z_i^A\}_{i=1}^{n-1}\big) \overset{(f)}{=} \frac{1}{k}\sum_{n=1}^{k}h(\tilde{z}_n|\tilde{z}_1^{n-1})$$

where (a) follows from the definition of β_k, (b) follows from (18), (c) holds because $2h(K\tilde{z}_n|a\tilde{e}_n) = \log(K^2) + 2h(\tilde{z}_n|a\tilde{e}_n)$ (see, e.g. [21, Th. 8.6.4]), (d) is due to the arithmetic mean and geometric mean inequality, (e) follows by applying the data processing inequality [21, Corollary 2.8.1] to the Markov chain $\{\hat{\tilde{x}}_n, \hat{x}_n^A\} \to \{z_i^A\}_{i=1}^{n-1} \to \tilde{z}_n$, and (f) follows from Assumption 2. ■

REFERENCES

[1] C.-Z. Bai and V. Gupta, "On Kalman filtering in the presence of a compromised sensor: Fundamental performance bounds," in *Proc. Amer. Control Conf.*, Portland, OR, USA, Jun. 2014, pp. 3029–3034.

[2] J. P. Farwell and R. Rohozinski, "Stuxnet and the future of cyber war," *Survival*, vol. 53, no. 1, pp. 23–40, 2011.

[3] G. Richards, "Hackers vs slackers," *Eng. Technol.*, vol. 3, no. 19, pp. 40–43, 2008.

[4] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on SCADA systems," *IEEE Trans. Control Syst. Technol.*, vol. 22, no. 4, pp. 1396–1407, Jul. 2014.

[5] A. A. Cárdenas, S. Amin, and S. S. Sastry, "Research challenges for the security of control systems," in *Proc. 3rd Conf. Hot Topics Security*, Berkeley, CA, USA, 2008, pp. 6:1–6:6.

[6] R. Patton, P. Frank, and R. Clark, *Fault Diagnosis in Dynamic Systems: Theory and Applications*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1989.

[7] M. Basseville and I. V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[8] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Trans. Automat. Control*, vol. 58, no. 11, pp. 2715–2729, Nov. 2013.

[9] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Trans. Automat. Control*, vol. 59, no. 6, pp. 1454–1467, Jun. 2014.

[10] G. Basile and G. Marro, *Controlled and Conditioned Invariants in Linear System Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1991.

[11] S. Sundaram and C. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of malicious agents," *IEEE Trans. Automat. Control*, vol. 56, no. 7, pp. 1495–1508, Jul. 2011.

[12] S. Cui, Z. Han, S. Kar, T. T. Kim, H. V. Poor, and A. Tajer, "Coordinated data-injection attack and detection in the smart grid: A detailed look at enriching detection solutions," *IEEE Signal Process. Mag.*, vol. 29, no. 5, pp. 106–115, Jul. 2012.

[13] C.-Z. Bai, F. Pasqualetti, and V. Gupta, "Security in stochastic control systems: Fundamental limitations and performance bounds," in *Proc. Amer. Control Conf.*, Chicago, IL, USA, Jul. 2015, pp. 195–200.

[14] S. Mishra, Y. Shoukry, N. Karamchandani, S. Diggavi, and P. Tabuada, "Secure state estimation: Optimal guarantees against sensor attacks in the presence of noise," in *Proc. Int. Symp. Inf. Theory*, Jun. 2015, pp. 2929–2933.

[15] Z. Guo, D. Shi, K. H. Johansson, and L. Shi, "Optimal linear cyber-attack on remote state estimation," *IEEE Trans. Control Netw. Syst.*, vol. 4, no. 1, pp. 4–13, Mar. 2016.

[16] E. Kung, S. Dey, and L. Shi, "The performance and limitations of $epsilon$-stealthy attacks on higher order systems," *IEEE Trans. Automat. Control*, vol. 62, no. 2, pp. 941–947, Feb. 2017.

[17] R. Zhang and P. Venkitasubramaniam, "Stealthy control signal attacks in vector lqg systems," in *Proc. Amer. Control Conf.*, Boston, MA, USA, 2016, pp. 1179–1184.

[18] S. Weerakkody, B. Sinopoli, S. Kar, and A. Datta, "Information flow for security in control systems," in *Proc. IEEE Conf. Decis. Control*, 2016, pp. 5065–5072.

[19] Y. Chen, S. Kar, and J. M. F. Moura, "Optimal attack strategies subject to detection constraints against cyber-physical systems," arXiv:1610.03370, 2016.

[20] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2000.

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[22] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Pacific Grove, CA, USA: Duxbury, 2002.

[23] S. Ihara, *Information Theory for Continuous Systems*, vol. 2. Singapore: World Sci., 1993.