2014 American Control Conference (ACC)
June 4-6, 2014. Portland, Oregon, USA

On Kalman Filtering in the Presence of a Compromised Sensor: Fundamental Performance Bounds

Cheng-Zong Bai and Vijay Gupta

Abstract— **Consider a scalar linear time-invariant system whose state is being estimated by an estimator using measurements from a single sensor. The sensor may be compromised by an attacker. The attacker is allowed to replace the measurement sequence by an arbitrary sequence. When the estimator uses this sequence, its estimate is degraded in the sense that the mean square error of this estimate is higher. The estimator monitors the received data to detect if an attack is in progress. The aim of the attacker is to degrade the estimate to the maximal possible amount while remaining undetected. By defining a suitable notion of stealthiness of the attacker, we characterize the trade-off between the fundamental limits of performance degradation that an attacker can induce versus its level of stealthiness. For various information patterns that characterize the information available at every time step to the attacker, we provide information theoretic bounds on the worst mean squared error of the state estimate that is possible and provide attacks that can achieve these bounds while allowing the attacker to remain stealthy even if the estimator uses arbitrary statistical ergodicity based tests on the received data.**

I. Introduction

Security of cyber-physical systems is now an established research topic. The canonical problem in this field is for an attacker to attack the cyber-system, i.e., the data being transmitted, to induce performance degradation in the physical system by corrupting the estimate or generate inappropriate control. Conversely, the problem for the estimator or the controller is to detect if the received data has been corrupted, and if so, to mitigate the performance loss due to such corrupted data.

Various attack models have been defined and developed in the literature. While simple attacks may consider an attacker that jams the channel across which data is transmitted, more sophisticated attacks that substitute false data for the correct measurements or control signals may achieve greater performance degradation. For any attacker, two metrics seem relevant: the performance degradation it can induce and the ease with which an attack it generates can be detected by the estimator or the controller. It seems intuitive that an attacker that does not care about being detected can substitute arbitrarily bad data and degrade the system performance arbitrarily. It is the constraint that the attack be undetected that limits the actions of the attacker and consequently the performance degradation it can induce. This trade-off has been studied in the literature. For deterministic systems, works such as [1], [2], [3], [4], [5] consider an attacker that can control some sensors and controllers in a distributed deterministic system. Since the system is deterministic, the only degree of freedom available to the attacker if she wants to remain stealthy is to utilize the fact that not all modes of the system can be observed by every sensor. These works relate stealthiness to the observability structure of the system and compute the performance degradation that such a stealthy attacker can induce. A different stream of work considers stochastic systems in which the attacker has an additional degree of freedom in the sense that the process and measurement noises induce some uncertainty in what the correct values of the measurements should be. Most works in this direction consider stealthiness of the attacker with respect to specific schemes employed by the estimator to detect if an attack is in progress. Usually such schemes follow from classical bad data detection algorithms [6] and are in the form of some residual error detector. Finally, there is a rich and growing literature of works (see, e.g., [1], [2], [3], [4], [7], [8], [9], [10], [11]) that define various specific attack strategies and detection tests for such attacks when the attacker's capability is constrained in some manner - say in terms of the number of components that it can compromise.

Thus, there seems to be a gap in the literature. While the notion of stealthiness for distributed deterministic systems has been defined, a similar notion for stochastic systems has been studied only for the special case when the detector employs a particular test, or when there are additional constraints on the attacker. In this work, we take the first step towards filling this gap. We focus on state estimation for an autonomous system (see, e.g., [1], [6], [12] for a good overview on works with this setup). We consider a process being driven by white noise whose state is observed by one sensor. An estimator relies on the measurement sequence it receives from the sensor to generate an estimate in the minimum mean squared error (MMSE) sense. The sensor may be compromised and the attacker can substitute any arbitrary sequence for the correct measurement sequence. The estimator has to detect if an attack is in progress. If an attack is not detected, the state estimate is calculated. Note that the system is centralized and fully observable; however, the stochastic measurement and process noises allow the attacker some stealthiness in altering the data.

For this problem, we begin by defining a suitable notion of stealthiness in terms of the probability of missed detection when an attack is in progress and probability of false alarm when no attack is in progress. The estimator must rely on the statistical properties of the received data sequence to detect an attack. We allow the estimator to conduct any ergodicity based test in which sample averages, moments

Cheng-Zong Bai and Vijay Gupta are with the Department of Electrical Engineering, University of Notre Dame, IN 46556 {cbai, vgupta2}@nd.edu

978-1-4799-3274-0/$31.00 ©2014 AACC

or distribution function can be checked against expected values. Using information theoretic tools, we characterize the asymptotic performance degradation that can be induced by an attacker that remains stealthy. For various information patterns that specify the information available to the attacker, we present explicit attacks that achieve this bound.

The paper is organized as follows. We begin in Section II by presenting the system and attacker model and formulating the problem. In Section III, we first present a converse result for the maximum performance degradation that a stealthy attacker can induce. Then, we present achievability results for various information patterns for the attacker. Section V presents numerical results to illustrate the trade-off between the performance degradation and stealthiness. Section VI concludes the paper.

II. System and Attack Model

Consider the scalar linear time-invariant system:

$$x_{k+1} = ax_k + w_k, \qquad y_k = cx_k + v_k \tag{1}$$

with the initial condition x_1 that is assumed to be a Gaussian random variable with mean zero. Without loss of generality, we assume that $c \geq 0$. In this system model, $\{w_k\}_{k>0}$ and $\{v_k\}_{k>0}$ represent the process noise and the measurement noise, respectively; both sequences are assumed to be independent and identically distributed (i.i.d.) Gaussian processes with mean zero and variance σ_w^2, σ_v^2, respectively. Moreover, we assume that the process noise and the measurement noise are mutually independent.

An estimator receives the measurements $\{y_n\}_{n=1}^k$ and generates a minimum mean squared error estimate of the state based on these measurements. Denote the estimate of x_{k+1} based on the measurement sequence $\{y_n\}_{n=1}^k$ by $\hat{x}_{k+1}$. Denote the corresponding estimation error by $e_{k+1} \triangleq \hat{x}_{k+1} - x_{k+1}$ and the mean square error (MSE) by $P_{k+1} \triangleq \mathbb{E}\big[e_{k+1}^2\big]$. It is well-known that the Kalman filter [13], [14] provides a recursive calculation for the estimate as

$$\hat{x}_{k+1} = a\hat{x}_k + K_k(y_k - c\hat{x}_k) \tag{2}$$

where the Kalman gain and the MSE P_{k+1} can be calculated by the following recursions:

$$K_k = \frac{acP_k}{c^2P_k + \sigma_v^2}, \quad P_{k+1} = a^2P_k + \sigma_w^2 - \frac{a^2c^2P_k^2}{c^2P_k + \sigma_v^2}.$$

The initial condition of the Kalman filter is given by $\hat{x}_1 = \mathbb{E}[x_1] = 0$. The sequence $\{z_k\}_{k>0}$ calculated as $z_k \triangleq y_k - c\hat{x}_k$ is called the innovation sequence. It is well known that the innovation sequence is a zero mean white Gaussian process with variance $\mathbb{E}[z_k^2] = c^2P_k + \sigma_v^2$.

If the system (1) is detectable (i.e., $|a| < 1$ or $c \neq 0$), then the Kalman filter converges to a steady state in the sense that $\lim_{k\to\infty} P_k$ exists [14]. Denote by $P \triangleq \lim_{k\to\infty} P_k$ the asymptotic MSE. The asymptotic MSE P is the positive semi-definite solution of the following algebraic Riccati equation:

$$P = a^2P + \sigma_w^2 - \frac{a^2c^2P^2}{c^2P + \sigma_v^2}, \tag{3}$$

and hence the steady-state Kalman gain $K \triangleq \lim_{k\to\infty} K_k$ is obtained by $K = \frac{acP}{c^2P+\sigma_v^2}$. Other statistical properties of the steady-state Kalman filter are given by $\sigma_z^2 \triangleq \lim_{k\to\infty} \mathbb{E}\big[z_k^2\big] = c^2P + \sigma_v^2$, $\sigma_x^2 \triangleq \lim_{k\to\infty} \mathbb{E}\big[x_k^2\big] = \frac{\sigma_w^2}{1-a^2}$, $\sigma_{\hat{x}}^2 \triangleq \lim_{k\to\infty} \mathbb{E}\big[\hat{x}_k^2\big] = \frac{K^2\sigma_z^2}{1-a^2}$.

We are interested in the situation when the sensor in the system (1) is compromised by an attacker who is capable of replacing the measurement sequence $\{y_k\}_{k>0}$ by any arbitrary attack sequence $\{\tilde{y}_k\}_{k>0}$. However, the attack sequence must be a function only of the information available at the attacker, as specified by the information pattern of the problem. If the estimator is not aware of the presence of the attacker, the attack sequence $\{\tilde{y}_k\}_{k>0}$ is treated as the input of the Kalman filter. Denote the corresponding "estimate" of the state obtained from the output of the Kalman filter by $\{\hat{\tilde{x}}_k\}_{k>0}$. Similar to (2), the sequence $\{\hat{\tilde{x}}_k\}_{k>0}$ is obtained by the recursion $\hat{\tilde{x}}_{k+1} = a\hat{\tilde{x}}_k + K_k\tilde{z}_k$, where the initial condition $\hat{\tilde{x}}_1 = \hat{x}_1$ is not changed. Denote the corresponding "innovation" by $\tilde{z}_k \triangleq \tilde{y}_k - c\hat{\tilde{x}}_k$. Note that the sequence $\{\tilde{z}_k\}_{k>0}$ need neither be zero mean, nor white or Gaussian.

For the estimation error $\tilde{e}_{k+1} = \hat{\tilde{x}}_{k+1} - x_{k+1}$ of this compromised estimate, denote the second moment by $\tilde{P}_{k+1} = \mathbb{E}[\tilde{e}_{k+1}^2]$. In general, $\tilde{P}_{k+1} \geq P_{k+1}$ and the attacker is interested in maximizing $\tilde{P}_{k+1}$. We consider the asymptotic behavior of $\tilde{P}_{k+1}$ as the metric of the performance degradation that the attacker can induce. Since the attack sequence is arbitrary, $\tilde{P}_{k+1}$ may not converge. Accordingly, we consider the limit superior of the Cesàro mean of the sequence $\{\tilde{P}_k\}_{k>0}$ as given by

$$\tilde{P} \triangleq \limsup_{k\to\infty} \frac{1}{k} \sum_{n=1}^{k} \tilde{P}_n. \tag{4}$$

Notice that if $\{\tilde{P}_k\}_{k>0}$ is a convergent sequence, then $\tilde{P} = \lim_{k\to\infty} \tilde{P}_k$ (see, e.g., [15]).

A. Information Pattern for Generating an Attack

The attack sequence is a function of the information available to the attacker. We denote by $\mathcal{I}_k$ the information available to the attacker at time k. Specification of $\mathcal{I}_k$ for every k indicates the information pattern for the attacker.

Assumption 1: The sequence $\{\mathcal{I}_k\}_{k>0}$ is assumed to have the following properties.

1) Due to causality constraints, $\mathcal{I}_k$ is independent of $\{w_n\}_{n\geq k}$ and $\{v_n\}_{n>k}$.
2) The attacker is assumed to have knowledge of the system parameters $a, c, \sigma_w^2, \sigma_v^2$.

B. Stealthiness

Intuitively, an attacker can degrade the performance of the estimator by an arbitrary amount if there is no constraint on the attack sequence $\{\tilde{y}_k\}_{k>0}$. For instance, it can simply set the measurements to very large values, leading to $\tilde{P}$ attaining an arbitrarily high value. In a practical setting, there are constraints on the attack sequence that arise from the attacker's desire to stay undetected. If the attacker is detected, all received measurements can be ignored and the

degradation of the MSE will be no greater than $\mathbb{E}[x_k^2]$, which is bounded for a stable system. Thus, the attacker aims at degrading the performance of the estimator by more than this amount while remaining undetected or stealthy.

To define a stealthy attack more formally, we separate the function of an attack detector from that of an estimator. To detect a stealthy attack, the attack detector must rely on the statistical properties of the received measurement sequence as compared with their expected values.

Definition 1 (Stealthy Attack): An attack $\{\tilde{y}_k\}_{k>0}$ is said to be stealthy if there exists no detector that can detect that an attack is in progress with zero probability of false alarm and zero probability of missed detection.

Since the initial condition of the Kalman filter is not changed by any attacker, there is a one-to-one correspondence between the measurement sequence $\{y_k\}_{k>0}$ (resp. $\{\tilde{y}_k\}_{k>0}$) and the innovation sequence $\{z_k\}_{k>0}$ (resp. $\{\tilde{z}_k\}_{k>0}$). Since the innovation sequence has a nice statistical characterization, it is convenient to use the innovation sequences $\{z_k\}_{k>0}$ and $\{\tilde{z}_k\}_{k>0}$ in the sequel. By the Neyman-Pearson criterion [16], if the two random sequences $\{z_k\}_{k>0}$ and $\{\tilde{z}_k\}_{k>0}$ are identically distributed, then the corresponding attack is stealthy. We provide two examples of such stealthy attacks and the corresponding performance degradation generated by attackers with access to different information patterns.

Example 1: Suppose that the attacker has no information about the states or the measurements. The attacker generates two random sequences $\{\tilde{w}_k\}_{k>0}$ and $\{\tilde{v}_k\}_{k>0}$ that are identically distributed to, but independent of, $\{w_k\}_{k>0}$ and $\{v_k\}_{k>0}$, respectively. Consider the attack generated by substituting the true measurements by the measurements that are the output of a fictitious system with the same dynamics as the true plant dynamics in (1) but the process noise and the measurement noise replaced by $\{\tilde{w}_k\}_{k>0}$ and $\{\tilde{v}_k\}_{k>0}$, respectively. Then, the attack $\{\tilde{y}_k\}_{k>0}$ is stealthy. Since $\hat{\tilde{x}}_{k+1}$ and x_{k+1} are independent, the MSE can be calculated as $\tilde{P}_{k+1} = \mathbb{E}\big[\hat{\tilde{x}}_{k+1}^2\big] + \mathbb{E}\big[x_{k+1}^2\big]$. Hence, we have

$$\tilde{P} = \begin{cases} \sigma_{\hat{x}}^2 + \sigma_x^2, & \text{if } |a| < 1, \\ \infty, & \text{otherwise.} \end{cases} \tag{5}$$

Example 2: Consider the attack $\tilde{y}_k = -y_k$ that can be generated by an attacker with the information $\mathcal{I}_k = \{y_n\}_{n=1}^k$. Obviously, we have $\tilde{z}_k = -z_k$, and the attack is stealthy. To calculate the MSE for the attack, notice that $\hat{\tilde{x}}_{k+1} = -\hat{x}_{k+1}$ because of the linearity of the Kalman filter. We have $\tilde{P}_{k+1} = P_{k+1} + 4\mathbb{E}\big[\hat{x}_{k+1}^2\big]$. Considering the limit of $\tilde{P}_{k+1}$ yields

$$\tilde{P} = \begin{cases} P + 4\sigma_{\hat{x}}^2, & \text{if } |a| < 1, \\ \infty, & \text{otherwise.} \end{cases} \tag{6}$$

Remark 1: As we can see in Example 1, for unstable systems ($|a| \geq 1$), it is possible for an attacker to generate an attack that leads to an arbitrarily bad estimate as $k \to \infty$, even if it has no information about the states. Moreover, such a performance degradation is achieved while the attacker is stealthy. Thus, we focus on the case when $|a| < 1$.

C. Problem Statement

It seems intuitive that the performance degradation that an attacker can induce depends both on the information available to it and whether it is stealthy or not. While various works (e.g., [1], [2], [4], [9], [11]) have considered particular detectors that the estimator can apply to detect if an attack is in progress, fundamental limits of the performance degradation that a stealthy attacker can induce when the estimator can apply any possible test is not known. Notice that the large body of literature (e.g. [3], [5], [10], [17]) that considers attacks that cannot be detected based on observability conditions are not directly applicable here, since the system is observable. As seen in Examples 1 and 2, in spite of the system being observable, an attacker can utilize the statistical uncertainty introduced by the process and measurement noises to induce a high $\tilde{P}$. In this paper, we wish to obtain the fundamental limits of performance degradation, as measured by $\tilde{P}$, that a stealthy attacker can induce when the detector can apply any ergodicity based test. To solve this problem, we use tools from information theory to bound the performance degradation that an attacker can induce if it wishes to remain stealthy.

III. Main Results

The problem statement as stated above is still quite general. To make inroads into the problem, we make the following further assumptions.

Assumption 2:

1) For a stable system, the initial condition of the Kalman filter is assumed to be $P_1 = P$.
2) The detector belongs to the class of ergodicity based detectors, i.e., any detector that uses $\frac{1}{k}\sum_{n=1}^{k} g(\zeta_n)$ as a test statistic for any function g that satisfies $\mathbb{E}[|g(z_n)|] < \infty$, where $\zeta_n = z_n$ if no attack is in progress, and $\zeta_n = \tilde{z}_n$ otherwise.

With the above assumption, the Kalman filter becomes time-invariant, namely, $K_k = K$ and $P_k = P$ for all $k \in \mathbb{N}$. Moreover, the innovation sequence $\{z_k\}_{k>0}$ becomes an i.i.d. sequence of Gaussian random variables with mean zero and variance σ_z^2. We will like to emphasize that this assumption is solely for ease of presentation and all the main results in the paper go through without this assumption by using the ergodic theory of non-stationary processes [18]. An ergodicity based detector as defined in Assumption 2 relies on the strong law of large numbers and can detect any attack sequence whose sample averages do not coincide with the marginally expected statistical averages. While such a detector is not the most general detector possible, it still represents a very powerful detector that can consider arbitrary statistics of the received data such as any moment of the data or its marginal distribution. We will also find it convenient to consider an analog notion of stealthiness rather than the digital notion implied by Definition 1. We define the following notion of stealthiness, which is a generalization of Definition 1.

Definition 2: (Marginal Stealthiness) An attack $\{\tilde{y}_k\}_{k>0}$ is said to be δ-marginally stealthy (δ-MS) if its associated

innovation sequence $\{\tilde{z}_k\}_{k>0}$ satisfies the condition

$$\mathbb{P}\left[\lim_{k\to\infty}\left\{\left|\frac{1}{k}\sum_{n=1}^{k} g(\tilde{z}_n) - \mathbb{E}\big[g(z_n)\big]\right| \le \delta\right\}\right] = 1 \qquad (7)$$

for any measurable function g where $\big|\mathbb{E}[g(z_n)]\big| \le 1$. Further, an attack is said to be strictly MS if it is δ-MS with $\delta = 0$.

Notice that stealthiness as defined in Definition 1 if specialized for the class of ergodicity based detectors that we consider here corresponds to the case when $\delta = 0$.

A. Preliminary Results

Our main contribution is a fundamental limit for the performance degradation induced by a δ-MS attack. We will provide a converse result and an achievability result for this limit. To prove these results, we begin by considering the performance limit for an even weaker notion of stealthiness.

Definition 3: (Weakly Marginal Stealthiness) An attack sequence $\{\tilde{y}_k\}_{k>0}$ is said to be ϵ-weakly marginally stealthy (ϵ-WMS) if its associated innovation sequence $\{\tilde{z}_n\}_{n=1}^{k}$ is ϵ-weak typical [15] almost surely (a.s.) as $k \to \infty$, namely,

$$\mathbb{P}\left[\lim_{k\to\infty}\left\{\left|\frac{1}{k}\sum_{n=1}^{k} -\log f_z(\tilde{z}_n) - h(z)\right| \le \epsilon\right\}\right] = 1 \qquad (8)$$

where $f_z(\cdot)$ and $h(z)$ are the probability density function and the differential entropy of z_n, respectively. Further, an attack is said to be strictly WMS if it is ϵ-WMS with $\epsilon = 0$.

The advantage of considering this notion is that an attack that is not ϵ-WMS is also not δ-MS for a suitably defined δ. Thus, the performance degradation induced by an ϵ-WMS provides an upper bound for the degradation that can be induced by a δ-MS attack. However, the notion of WMS attacks allows us to use the tool of weak typicality from information theory. Given the Gaussianity assumptions in the model (1), the following result is apparent.

Lemma 1: If an attack $\{\tilde{y}_k\}_{k>0}$ is ϵ-WMS, then

$$\limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\big[\tilde{z}_n^2\big] \le \limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\tilde{z}_n^2 \le (1+2\epsilon)\sigma_z^2 \quad (9)$$

Proof: This two inequalities follow immediately from Fatou's lemma [19] and (8), respectively. ∎

We begin by deriving the converse result for ϵ-WMS attacks by providing a preliminary result.

Lemma 2: For any random variables x and y, and any measurable function g, we have $\mathbb{E}\big[xg(y)\big] \le \mathbb{E}\big[\mathbb{E}[x|y]^2\big]^{\frac{1}{2}} \times \mathbb{E}\big[g(y)^2\big]^{\frac{1}{2}}$. Further, the inequality holds with equality if and only if $g(y)$ is chosen such that $\mathbb{E}[x|y] = \alpha g(y)$ almost surely where α is any non-negative constant.

Proof: The lemma follows from Cauchy-Schwarz inequality. ∎

Now we present the converse part of the performance degradation limit for ϵ-WMS attacks.

Theorem 1: Consider system (1) with the additional assumptions that $0 < |a| < 1$ and $c > 0$. Suppose that the attacking sequence $\{\tilde{y}_k\}_{k>0}$ generated by an attacker with access to information sets $\{\mathcal{I}_k\}_{k>0}$ is ϵ-WMS. Then the limit superior of the Cesàro mean of the MSE is bounded by

$$\tilde{P} \le \frac{1}{1-a^2}\Big(\big(K^2 + 2aK\alpha\big)(1+2\epsilon)\sigma_z^2 + \sigma_w^2\Big), \qquad (10)$$

where $\alpha > 0$ is chosen such that all the inequalities in (1) hold with equalities if we let $\mathbb{E}\big[\tilde{e}_n\big|\mathcal{I}_n\big] = \alpha\tilde{z}_n$ for every $n \in \mathbb{N}$.

Proof: For every ϵ-WMS attack, its associated innovation sequence satisfies the inequality in Lemma 1. To obtain an upper bound of the Cesàro mean of the MSE, we consider a superset of ϵ-WMS attacks, that is, the set of attacks whose associated innovation sequence satisfies the inequality in Lemma 1. Note that w_n is independent of $\{x_i\}_{i=1}^{n}$ and $\mathcal{I}_n$ because of the causality assumption. Thus, we obtain

$$\begin{aligned}
&\limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\big[\tilde{e}_{n+1}^2\big] \\
&= \limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k} a^2\tilde{P}_n + K^2\mathbb{E}\big[\tilde{z}_n^2\big] + 2aK\mathbb{E}\big[\tilde{e}_n\tilde{z}_n\big] + \sigma_w^2 \\
&\le a^2\limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\tilde{P}_n + \limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k} K^2\mathbb{E}\big[\tilde{z}_n^2\big] \\
&\quad + \limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k} 2aK\mathbb{E}\big[\tilde{e}_n\tilde{z}_n\big] + \sigma_w^2. \qquad (11)
\end{aligned}$$

Since $\hat{x}_1$ is the initial condition of the Kalman filter, which can not be altered by any attacker, we have $\tilde{P}_1 = P < \infty$. Hence, we can write $\limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\tilde{P}_{n+1} = \limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\tilde{P}_n$. From (9) and (11), we obtain

$$\tilde{P} \le \frac{K^2(1+2\epsilon)\sigma_z^2 + \limsup_{k\to\infty}\frac{2aK}{k}\sum_{n=1}^{k}\mathbb{E}\big[\tilde{e}_n\tilde{z}_n\big] + \sigma_w^2}{1-a^2}. \qquad (12)$$

Due to the assumption of $c > 0$, we have $aK \ge 0$. In order to maximize (12), we have to find an upper bound for $\sum_{n=1}^{k}\mathbb{E}[\tilde{e}_n\tilde{z}_n]$ by choosing an appropriate $\tilde{z}_n$ as a function of $\mathcal{I}_n$ for $n = 1, 2, \cdots, k$. Using Lemma 2 and Cauchy-Schwarz inequality yields

$$\begin{aligned}
\frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\big[\tilde{e}_n\tilde{z}_n\big] &\le \frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\Big[\mathbb{E}\big[\tilde{e}_n\big|\mathcal{I}_n\big]^2\Big]^{\frac{1}{2}}\mathbb{E}\big[\tilde{z}_n^2\big]^{\frac{1}{2}} \qquad (13)\\
&\le \left(\frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\Big[\mathbb{E}\big[\tilde{e}_n\big|\mathcal{I}_n\big]^2\Big]\right)^{\frac{1}{2}}\left(\frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\big[\tilde{z}_n^2\big]\right)^{\frac{1}{2}}. \qquad (14)
\end{aligned}$$

By Lemma 2, the inequality (13) holds with equality if and only if $\mathbb{E}\big[\tilde{e}_n\big|\mathcal{I}_n\big] = \alpha_n\tilde{z}_n$ holds for every $n \in \{1, 2, \cdots, k\}$ where $\alpha_n \ge 0$. Both inequalities (13) and (14) hold with equalities if and only if $\alpha_1 = \cdots = \alpha_k = \alpha$. Hence,

$$\limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\mathbb{E}\big[\tilde{e}_n\tilde{z}_n\big] \le \alpha(1+2\epsilon)\sigma_z^2 \qquad (15)$$

where α is chosen as indicated in the statement of the theorem. By (12) and (15), the theorem follows. ∎

Remark 2:

- Note that the upper bound in (10) may not be achievable, since this upper bound is obtained by considering a superset of ϵ-WMS attacks.
- The upper bound in (10) is a monotonic increasing function of ϵ. From the attacker's perspective, this introduces a trade-off between the degree of stealthiness (as measured by the value of ϵ) and the performance degradation it can induce at the estimator.

Theorem 1 provides a converse statement that no ϵ-WMS attack can cause the Cesàro mean of the MSE to be larger than the upper bound in (10). The upper bound in (10) is in terms of the parameter α that is a function of the specific information available to the attacker. We now prove the achievablity of (10) for several information patterns.

Example 3 ($\mathcal{I}_k = \{\tilde{z}_n\}_{n=1}^{k-1}$): Suppose that the attacker has no information about the states. However, it is reasonable to assume that the attacker knows the attack sequence it generated in the past. To achieve the upper bound in Theorem 1, the attack is generated by the relation $\mathbb{E}[\hat{\tilde{x}}_k - x_k | \{\tilde{z}_n\}_{n=1}^{k-1}] = \alpha \tilde{z}_k$, which gives a recursion $\alpha \tilde{z}_{k+1} = \big(a\alpha + K\big)\tilde{z}_k$. The innovation sequence associated with the attack is given by $\tilde{z}_k = (-1)^{k+1}\tilde{z}_1$ if $-1 < a < 0$ and $\tilde{z}_k = \tilde{z}_1$ if $0 < a < 1$. In addition, the choice of initial condition is given by $\tilde{z}_1 = \pm\sqrt{(1+2\epsilon)}\sigma_z$. Clearly, the attack is ϵ-WMS. Thus the upper bound in Theorem 1 is achievable by an ϵ-WMS attack.

Next, we consider the information pattern $\{\mathcal{I}_k\}_{k>0}$ that enables the attacker to use a steady state Kalman filter to estimate the state x_{k+1} at every time k. Specifically, the attacker's MMSE state estimate can be written as

$$\hat{x}_{k+1}^A = a\hat{x}_k^A + K^A z_k^A, \tag{16}$$

where $K^A > 0$ is the attacker's Kalman gain and $\{z_k^A\}_{k>0}$ is the attacker's innovation sequence that is assumed to be i.i.d. Gaussian with mean zero and variance $\sigma_{z^A}^2 > 0$. We present three examples for such information pattern.

Example 4 ($\mathcal{I}_k = \{x_n\}_{n=1}^k$): In this case, $\hat{x}_{k+1}^A = ax_k$. Note that the attacker can recover the process noise w_{k-1} since $w_{k-1} = x_k - ax_{k-1}$. Such state estimate is compatible with (16) if we let $K^A = a$ and $z_k^A = w_{k-1}$.

Example 5 ($\mathcal{I}_k = \{y_n\}_{n=1}^{k-1}$): Consider the case in which the attacker can intercept $\{y_n\}_{n=1}^{k-1}$ and then generate $\tilde{y}_k$. In this case, $\hat{x}_{k+1}^A = a\hat{x}_k$, and it can be obtained by letting $K^A = aK$ and $z_k^A = z_{k-1}$.

Example 6 ($\mathcal{I}_k = \{y_n\}_{n=1}^k$): Clearly, we have $\hat{x}_{k+1}^A = \hat{x}_{k+1}$, and hence $K^A = K$, $z_k^A = z_k$.

For any such information pattern, let us consider the innovation sequence $\{\tilde{z}_k\}_{k>0}$ associated with the attack that satisfies $\mathbb{E}\big[\hat{\tilde{x}}_k - x_k \big| \mathcal{I}_k\big] = \alpha\tilde{z}_k$ for all $k \in \mathbb{N}$. Note that

$$\begin{aligned}
\alpha\tilde{z}_{k+1} &= (a\hat{\tilde{x}}_k + K\tilde{z}_k) - \Big(a\hat{x}_k^A + K^A z_k^A + \frac{K^A}{a} z_{k+1}^A\Big) \\
&= a\Big(\hat{\tilde{x}}_k - \big(\hat{x}_k^A + \frac{K^A}{a} z_k^A\big)\Big) + K\tilde{z}_k - \frac{K^A}{a} z_{k+1}^A \\
&= a\alpha\tilde{z}_k + K\tilde{z}_k - \frac{K^A}{a} z_{k+1}^A
\end{aligned}$$

since $\mathbb{E}\big[x_k \big| \mathcal{I}_k\big] = \hat{x}_k^A + \frac{1}{a}K^A z_k^A$ for every $k \in \mathbb{N}$. Hence, we obtain a recursion for generating $\{\tilde{z}_k\}_{k>0}$:

$$\tilde{z}_{k+1} = \Big(a + \frac{K}{\alpha}\Big)\tilde{z}_k - \frac{K^A}{a\alpha} z_k^A. \tag{17}$$

In particular, from (17), we have

$$\lim_{n\to\infty} \mathbb{E}\big[\tilde{z}_n^2\big] = \frac{(\frac{K^A}{a\alpha})^2 \sigma_{z^A}^2}{1 - \big(a + \frac{K}{\alpha}\big)^2} \tag{18}$$

if $\big|a + \frac{K}{\alpha}\big| < 1$. Now, we focus on the calculation of the constant α. We choose α such that $\lim_{k\to\infty} \mathbb{E}\big[\tilde{z}_k^2\big] = (1 + 2\epsilon)\sigma_z^2$. Using (18), α can be solved by

$$\alpha = \frac{aK + \sqrt{K^2 + \big(\frac{K^A}{a}\big)^2 \cdot \frac{(1-a^2)\sigma_{z^A}^2}{(1+2\epsilon)\sigma_z^2}}}{1 - a^2}. \tag{19}$$

Notice that such an attack makes all the inequalities in the proof of Theorem 1 hold with equality. Moreover, $\{\tilde{z}_k\}_{k>0}$ is an autoregressive process of order one. This implies that $\{\tilde{z}_k\}_{k>0}$ is ergodic [20], which further implies that all the inequalities in (9) hold with equality and such an attack is ϵ-WMS. Thus, the upper bound in Theorem 1 is achievable.

B. The Fundamental Limit for MS Attacks

Using the converse and the achievability of ϵ-WMS attacks, we can now present the main result, which is the fundamental limit of performance degradation induced by δ-MS attacks.

Theorem 2: Consider the problem formulation presented in Section II. Suppose that $0 < |a| < 1$ and $c > 0$.

1) (Converse) The the upper bound (10) holds for all δ-MS attacks, where ϵ is chosen as
$$\epsilon = \begin{cases} \delta \quad , & \text{if } |h(z)| \le 1, \\ \frac{\delta}{|h(z)|} \, , & \text{otherwise.} \end{cases} \tag{20}$$
2) (Achievability) For the information patterns that possess the property stated in (16), the upper bound (10) is achievable.

Proof: 1) With the ϵ stated in (20), clearly, δ-MS implies ϵ-WMS. Using Theorem 1, we can obtain the converse. 2) Since the attack generated according to (17) is ergodic, the performance bound in (10) is achievable. ■

Remark 3:

1) We would like to emphasize that the achievability result in Theorem 2 holds only for $\delta = 0$.
2) Since strict MS is implied by the stealthiness in Definition 1, $\tilde{P}$ achieved by any stealthy attack is also upper bounded by (10) with $\epsilon = 0$.
3) The optimal strict MS attacks for the information patterns in Theorem 2 are not stealthy in the sense of Definition 1, since such attacks can be detected by estimating the autocorrelation of $\{\tilde{z}_k\}_{k>0}$.

IV. Numerical Results

So far, we have presented six different attacks in Examples 1–6. We denote by $\tilde{P}_{\#i}$ the $\tilde{P}$ achieved by the attack in Example i where $i \in \{1, 2, \cdots, 6\}$. Unless otherwise specified, in this section, we let $(a, c, \sigma_w^2, \sigma_v^2) = (0.4, 1, 1, 0.1)$.

A. The MSE Due to the Attacks

For the attacks in Examples 3–6, we set $\epsilon = 0$. We discuss the effect on the ratio $P/\tilde{P}$ by changing the value of a, as shown in Fig. 1. The parameter a measures the correlation of the system state at time $k+1$ to the state at time k. Intuitively, $a = 0$ would imply that the system state x_{k+1} is simply the noise w_k. Thus, an attacker can not degrade the MSE any further. Figure 1 confirms this intuition and further presents the degradation for various values of a, as attacks with different information patterns are considered. Figure 1 also shows that, as $a \to 0$, the attack in Example 2 is the optimal stealthy attack for the information patterns $\mathcal{I}_k = \{x_n\}_{n=1}^k$ and $\mathcal{I}_k = \{y_n\}_{n=1}^k$, since the set of stealthy attacks is a subset of all strict MS attacks. Moreover, in this regime, the knowledge of $\{x_n\}_{n=1}^k$ is redundant as long as the attacker knows $\{y_n\}_{n=1}^k$. On the other hand, as $|a| \to 1$, we can see that $\tilde{P}$ is unbounded for all attacks. Since the information patterns in Examples 3 and 4 indicate the two extreme cases that the attacker has no and full information about the causal states, respectively, it shows that using any information pattern to generate the optimal strict MS attack leads to the same $\tilde{P}$ as $|a| \to 1$.

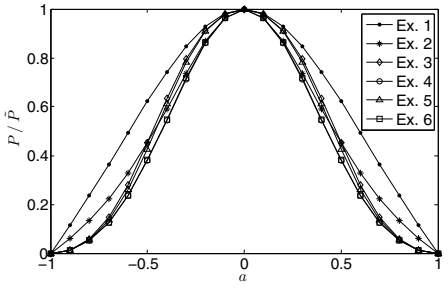

Fig. 1. $P/\tilde{P}$ v.s. a for several attacks where $c = 1, \sigma_w^2 = 1, \sigma_v^2 = 0.1$.

B. The Trade-off Between Stealthiness and MSE

Theorems 1 and 2 have formalized the notion that stealthiness of an attacker can be traded-off with the performance degradation it can induce. To illustrate such a trade-off numerically, we plot in Figure 2 the upper bound in (10) as a function of ϵ for the information pattern discussed above. Observe that, the slopes of the upper bounds in Figure 2 do not seem to be sensitive to the particular information pattern used. Indeed, it can be proved analytically that the limits of α in these examples go to infinity with the same rate. Consequently, as ϵ being large, the upper bounds of $\tilde{P}$ in (10) for all information patterns are asymptotically equivalent.

V. Concluding Remarks

For an unstable system, it is possible for an attacker to make the MSE arbitrary bad. For a stable system, Theorem 1 quantifies a trade-off between stealthiness and attack performance. Further, we prove the achievability of fundamental limits of performance degradation induced by ϵ-WMS attacks and strict MS attacks for several information patterns. Under certain limit conditions, we show that the attack in Example 2 is also the optimal stealthy attack.

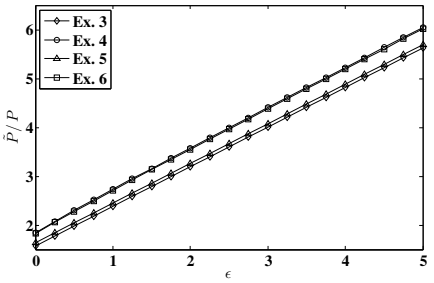

Fig. 2. The upper bound of $\tilde{P}/P$ v.s. ϵ for several attacks.

References

[1] A. Teixeira, D. Pérez, H. Sandberg, and K. H. Johansson, "Attack models and scenarios for networked control systems," in *Proc. of the 1st international conference on High Confidence Networked Systems*. ACM, 2012, pp. 55–64.

[2] Y. Mo and B. Sinopoli, "Secure control against replay attacks," in *47th Annual Allerton Conference*. IEEE, 2009, pp. 911–918.

[3] F. Pasqualetti, F. Dörfler, and F. Bullo, "Attack detection and identification in cyber-physical systems–part 1: Models and fundamental limitations," *arXiv preprint arXiv:1202.6144*, 2012.

[4] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," *ACM Trans. on Information and System Security*, vol. 14, no. 1, pp. 13, 2011.

[5] G. Dán and H. Sandberg, "Stealth attacks and protection schemes for state estimators in power systems," in *Smart Grid Communications, 2010 First IEEE International Conference on*. IEEE, 2010, pp. 214–219.

[6] 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," *Signal Processing Magazine, IEEE*, vol. 29, no. 5, pp. 106–115, 2012.

[7] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure state-estimation for dynamical systems under active adversaries," in *the 49th Annual Allerton Conference*. IEEE, 2011, pp. 337–344.

[8] H. Fawzi, P. Tabuada, and S. Diggavi, "Security for control systems under sensor and actuator attacks," in *IEEE 51st Annual Conference on Decision and Control (CDC)*. IEEE, 2012, pp. 3412–3417.

[9] O. Kosut, L. Jia, R. J. Thomas, and L. Tong, "Malicious data attacks on the smart grid," *Smart Grid, IEEE Trans. on*, vol. 2, no. 4, pp. 645–658, 2011.

[10] A. Teixeira, S. Amin, H. Sandberg, K. H. Johansson, and S.S. Sastry, "Cyber security analysis of state estimators in electric power systems," in *Decision and Control (CDC), 2010 49th IEEE Conference on*. IEEE, 2010, pp. 5991–5998.

[11] C. Kwon, W. Liu, and I. Hwang, "Security analysis for cyber-physical systems against stealthy deception attacks," in *American Control Conference (ACC), 2013*. IEEE, 2013, pp. 3344–3349.

[12] A. A. Cárdenas, S. Amin, Z.-S. Lin, Y-.L. Huang, C.-Y. Huang, and S. Sastry, "Attacks against process control systems: risk assessment, detection, and response," in *the 6th ACM Symp. on Information, Computer and Communications Security*. ACM, 2011, pp. 355–366.

[13] R. E. Kalman, "A new approach to linear filtering and prediction problems," *J. of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[14] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear Estimation*, Prentice Hall, 2000.

[15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 2nd edition, 2006.

[16] H. V. Poor, *An introduction to signal detection and estimation*, Springer-Verlag, New York, 2nd edition, 1998.

[17] S. D. Bopardikar and A. Speranzon, "On analysis and design of stealth-resilient control systems," in *IEEE Symposium on Resilient Control Systems*, 2013.

[18] R. M. Gray, *Probability, random processes, and ergodic properties*, Springer, 2009.

[19] P. Billingsley, *Probability and Measure*, Wiley, 2012.

[20] M. Taniguchi and Y. Kakizawa, *Asymptotic theory of statistical inference for time series*, Springer, New York, 2000.