Security in Stochastic Control Systems: Fundamental Limitations and Performance Bounds

Cheng-Zong Bai, Fabio Pasqualetti, and Vijay Gupta

Abstract— **This work proposes a novel metric to characterize the resilience of stochastic cyber-physical systems to attacks and faults. Specifically, we consider a single-input single-output plant regulated by a control law based on the estimate of a Kalman filter. We allow for the presence of an attacker able to hijack and replace the control signal. The objective of the attacker is to maximize the estimation error of the Kalman filter – which in turn quantifies the degradation of the control performance – by tampering with the control input, while remaining undetected. We introduce a notion of ϵ-stealthiness to quantify the difficulty to detect an attack when an *arbitrary* detection algorithm is implemented by the controller. For a desired value of ϵ-stealthiness, we quantify the largest estimation error that an attacker can induce, and we analytically characterize an optimal attack strategy. Because our bounds are independent of the detection mechanism implemented by the controller, our information-theoretic analysis characterizes fundamental security limitations of stochastic cyber-physical systems.**

I. Introduction

Cyber-physical systems offer a variety of attack surfaces arising from the interconnection of different technologies and components. Depending on their resources and capabilities, attackers generally aim to deteriorate the functionality of the system, while avoiding detection for as long as possible [1].

Security of cyber-physical systems is a growing research area where, recently, different attack strategies and defense mechanisms have been characterized. While simple attacks have a straightforward implementation and impact, such as jamming control and communication channels [2], sophisticated ones may degrade the functionality of a system more severely [3], [4], and are more difficult to mitigate. In this work we measure the severity of attacks based on their effect on the control performance and on their level of stealthiness, that is, the difficulty of being detected from measurements. Intuitively, there exists a trade-off between the degradation of control performance and the level of stealthiness of an attack. Although this trade-off has previously been identified for specific systems and detection mechanisms [5], [6], [7], [8], a thorough analysis of the resilience of stochastic control systems to arbitrary attacks is still missing.

Related works For deterministic cyber-physical systems the concept of stealthiness of an attack is closely related to the control-theoretic notion of zero dynamics [9]. In particular, an attack is undetectable if and only if it excites only the zero dynamics of an appropriately defined input-output system describing the system dynamics, the measurements available to a security monitors, and the variables compromised by the attacker [10], [11]. Thus, the question of stealthiness of an attack has a binary answer in deterministic systems. For stochastic cyber-physical systems, instead, the presence of process and measurements noise offers a smart attacker the additional possibility to tamper with sensor measurements and control inputs within the acceptable uncertainty levels, thereby making the detection task arbitrary difficult.

Detectability of attacks in stochastic systems has received only initial attention from the research community, and there seem to be no agreement on an appropriate notion of stealthiness. Most works in this area considers detectability of attacks with respect to specific detection schemes, such as the classic bad data detection algorithm [12]. In our previous work [13], we proposed the notion of ϵ-marginal stealthiness to quantify the stealthiness level with respect to the class of ergodic detectors. However, the notion of marginal stealthiness defined in [13] is still restricted in the class of ergodic detectors. Further, the detectability of an ϵ-marginally stealthy attack is not explicitly characterized in [13]. In this work we introduce a novel notion of stealthiness that is independent of the attack detection algorithm, and thus provides a fundamental measure of the stealthiness of attacks in stochastic control systems.

Contributions The contributions of this paper are threefold. First, we propose the notion of ϵ-stealthiness to quantify detectability of attacks in stochastic cyber-physical systems. Our metric is motivated by the Chernoff-Stein Lemma in detection and information theories [14], and is universal, in the sense that it is independent of any specific detection mechanism employed by the controller. Second, we provide an achievable bound for the degradation of the minimum-mean-square estimation error caused by an ϵ-stealthy attack, as a function of the system parameters, noise statistics, and information available to the attacker. Third and finally, we provide a closed-form expression of optimal ϵ-stealthy attacks achieving the maximal degradation of the estimation error. These results characterize the trade-off between performance degradation that an attacker can induce, versus the fundamental limit of the detectability of the attack.

We focus on single-input single-output systems with an observer-based controller. However, our methods are general, and applicable to multiple-input multiple-output systems via a more involved technical analysis.

Paper organization Section II contains our mathematical formulation of the problem and our model of attacker. In

This material is based upon work supported in part by awards NSF ECCS-1405330 and ONR N00014-14-1-0816. Cheng-Zong Bai and Vijay Gupta are with the Department of Electrical Engineering, University of Notre Dame, IN 46556, {cbai, vgupta2}@nd.edu . Fabio Pasqualetti is with the Department of Mechanical Engineering, University of California, Riverside, CA 92521, fabiopas@engr.ucr.edu .

Section III we discuss our metric to quantify the stealthiness level of an attack. The main results of this paper are presented in Section IV, including a characterization of the largest perturbation caused by an ϵ-stealthy attack, and a closed-form expression of optimal ϵ-stealthy attacks. Section V contains our illustrative examples and numerical results. Finally, Section VI concludes the paper.

Notation A sequence $\{x_n\}_{n=i}^{j}$ is denoted by x_i^j. A Gaussian random variable x with mean μ and variance σ^2 is denoted by $x \sim \mathcal{N}(\mu, \sigma^2)$.

II. System and Attack Models

A. System model

We consider the single-input single-output time-invariant system described by

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

where $a, c \in \mathbb{R}$, $c \neq 0$, w_1^∞ and v_1^∞ are random sequences representing process and measurement noise, respectively. We assume the sequences w_1^∞ and v_1^∞ to be independent and identically distributed (i.i.d.) Gaussian processes with $w_k \sim \mathcal{N}(0, \sigma_w^2)$, $v_k \sim \mathcal{N}(0, \sigma_v^2)$ for all $k > 0$. The control input u_k is generated based on a causal observer-based control policy, that is, u_k is a function of the measurement sequence y_1^k. In particular, the controller employs a Kalman filter [15], [16] to compute the Minimum-Mean-Squared-Error (MMSE) estimate $\hat{x}_{k+1}$ of x_{k+1} from the measurements y_1^k. The Kalman filter reads as

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

where the Kalman gain K_k and the MSE $P_{k+1} \triangleq \mathbb{E}\big[(\hat{x}_{k+1} - x_{k+1})^2\big]$ can be calculated by the recursions

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

with the initial condition $\hat{x}_1 = \mathbb{E}[x_1] = 0$ and $P_1 = \mathbb{E}[x_1^2]$. If the system (1) is detectable (i.e., $|a| < 1$ or $c \neq 0$), then the Kalman filter converges to the steady state in the sense that $\lim_{k\to\infty} P_k = P$ exists [16] where P can be obtained uniquely through the algebraic Riccati equation. For ease of presentation, we assume that $P_1 = P$. Hence, we obtain a steady state Kalman filter with Kalman gain $K_k = K$ and $P_k = P$ at every time step k. The sequence z_1^∞ calculated as $z_k \triangleq y_k - c\hat{x}_k$ is called the innovation sequence. Since we consider steady state Kalman filtering, the innovation sequence is an i.i.d. Gaussian process with $z_k \sim \mathcal{N}(0, c^2P + \sigma_v^2)$.

B. Attack model

We consider an attacker capable of hijacking and replacing the control input u_1^∞ with an arbitrary signal $\tilde{u}_1^\infty$. Assume that the attacker knows the system parameters $a, c, \sigma_w^2, \sigma_v^2$. The attack input $\tilde{u}_1^\infty$ is constructed based on the system parameters and the attacker *information pattern*. Let $\mathcal{I}_k$ denote the information available to the attacker at time k. We make the following assumptions on the attacker information patters:

(A1) the attacker knows the control input u_k, that is, $u_k \in \mathcal{I}_k$ at all times k;

(A2) the information available to the attacker is non-decreasing, that is, $\mathcal{I}_k \subseteq \mathcal{I}_{k+1}$;

(A3) $\mathcal{I}_k$ is independent of the w_k^∞ and v_{k+1}^∞ due to causality.

Attack scenarios satisfying assumptions (A1)–(A3) include:

(i) the attacker knows the control input at time k, that is, $\mathcal{I}_k = \{u_1^k\}$.

(ii) the attacker knows the control input and the state at time k, i.e., $\mathcal{I}_k = \{u_1^k, x_1^k\}$.

(iii) the attacker knows the control input and the (delayed) measurements received by the controller at time k, that is, $\mathcal{I}_k = \{u_1^k, \tilde{y}_1^{k-d}\}$ with $d \geq 0$.

(iv) the attacker knows the control input and take additional measurements $\bar{y}_k$ at time k, that is, $\mathcal{I}_k = \{u_1^k, \bar{y}_1^k\}$.

Let $\tilde{y}_1^\infty$ be the sequence of measurements received by the controller in the presence of the attack $\tilde{u}_1^\infty$. Then, $\tilde{y}_1^\infty$ is generated by the system dynamics

$$x_{k+1} = ax_k + \tilde{u}_k + w_k, \qquad \tilde{y}_k = cx_k + v_k. \tag{3}$$

Notice that, because the controller is unaware of the attack, the corrupted measurements $\tilde{y}_1^\infty$, and hence the attack input $\tilde{u}_1^\infty$, drive the Kalman filter (2) as an external input. Let $\hat{\tilde{x}}_1^\infty$ be the estimate of the Kalman filter (2) in the presence of the attack $\tilde{u}_1^\infty$, which is obtained from the recursion

$$\hat{\tilde{x}}_{k+1} = a\hat{\tilde{x}}_k + K\tilde{z}_k + u_k,$$

where the innovation is $\tilde{z}_k \triangleq \tilde{y}_k - c\hat{\tilde{x}}_k$. Notice that (i) the estimate $\hat{\tilde{x}}_{k+1}$ is sub-optimal, because it is obtained by assuming the nominal control input, whereas the system is driven by the attack input, and (ii) the random sequence $\tilde{z}_1^\infty$ need neither be zero mean, nor white or Gaussian, because the attack input is arbitrary.

Let $\tilde{P}_{k+1} = \mathbb{E}[(\hat{\tilde{x}}_{k+1} - x_{k+1})^2]$ be the second moment of the estimation error $\hat{\tilde{x}}_{k+1} - x_{k+1}$, and assume that the attacker aims to maximize $\tilde{P}_{k+1}$. We consider the asymptotic behavior of $\tilde{P}_{k+1}$ to measure the performance degradation induced by the attacker. Since the attack sequence is arbitrary, the sequence $\tilde{P}_1^\infty$ may not converge. Accordingly, we consider the limit superior of arithmetic mean of the sequence $\tilde{P}_1^\infty$ as given by

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

Notice that if the sequence $\tilde{P}_1^\infty$ is convergent, then $\lim_{k\to\infty} \tilde{P}_{k+1} = \tilde{P}$, which equals the Cesàro mean[1] [14].

III. The Notion of Stealthiness for Stochastic Systems

In this section we motivate and define our notion of ε-stealthiness of attacks. We begin with the standard definition

[1] The steady state assumption is made in order to obtain an i.i.d. innovation sequence. If the Kalman filter starts from an arbitrary initial condition P_1, then the innovation sequence is an independent, asymptotically identically distributed, Gaussian process. This identity guarantees that the results for non-steady state Kalman filter coincide with the main result in this paper.

of Kullback-Leibler divergence (or relative entropy) [14], [17], which plays an important role in our notion of stealthiness.

Definition* 1: *(Kullback-Leibler divergence) Let x_1^k and y_1^k be two random sequences with joint probability density functions (pdf) $f_{x_1^k}$ and $f_{y_1^k}$, respectively. The Kullback-Leibler Divergence (KLD) between x_1^k and y_1^k equals

$$D\big(x_1^k\big\|y_1^k\big) = \int_{-\infty}^{\infty} \log \frac{f_{x_1^k}(\xi_1^k)}{f_{y_1^k}(\xi_1^k)} f_{x_1^k}(\xi_1^k) d\xi_1^k. \tag{4}$$

□

The KLD is a non-negative quantity that gauges the dissimilarity between two probability density functions. It should be observed that $D\big(x_1^k\big\|y_1^k\big) = 0$ if $f_{x_1^k} = f_{y_1^k}$. Also, the KLD is generally not symmetric, that is, $D\big(x_1^k\big\|y_1^k\big) \neq D\big(y_1^k\big\|x_1^k\big)$.

Notice that the system (3) with $\sigma_w^2 = 0$ and $\sigma_v^2 = 0$ (i.e., deterministic single-input single-output system) features no zero dynamics. Hence, every attack would be detectable [10]. However, the stochastic nature of the system provides an additional degree of freedom to the attacker, because the process noise and the measurement noise induce some uncertainty in the measurements. Building on this idea, we now formally define attack stealthiness. Consider the problem of detecting an attack from measurements. Note that the detector must rely on the statistical properties of the received measurement sequence as compared with their expected model in (1). This can be formulated by the following binary hypothesis testing problem:

H_0 : No attack is in progress (the controller receives y_1^k);
H_1 : Attack is in progress (the controller receives $\tilde{y}_1^k$).

Suppose that a detector is employed by the controller. Let p_k^F be the probability of false alarm (decide H_1 when H_0 is true) at time k and let p_k^D be the probability of detection (decide H_1 when H_1 is true) at time k. In detection theory, the performance of the detector can be characterized by the trade-off between p_k^F and p_k^D, namely, the Receiver Operating Characteristic (ROC) [18]. From the ROC perspective, the attack that is hardest to detect is the one for which, at every time k, there exists no detector that performs better than a random guess (e.g., to make a decision by flipping a coin) independent of the hypothesis. If a detector makes a decision via a random guess independent of the hypothesis, then the operating point of the ROC satisfies $p_k^F = p_k^D$. This motivates us to define a binary notion of stealthiness as follows.

Definition* 2: *(Strict stealthiness) An attack $\tilde{u}_1^\infty$ is said to be strictly stealthy if there exists no detector such that $p_k^F < p_k^D$ for any $k > 0$. □

Remark* 1: *(Strictly stealthy attack) Using Neyman-Pearson Lemma [18], an attack $\tilde{u}_1^\infty$ is strictly stealthy if and only if $D\big(\tilde{y}_1^k\big\|y_1^k\big) = 0$ for all $k > 0$. □

The reader may argue that strict stealthiness is a too restrictive notion of stealthiness for an attacker, and it significantly limits the set of stealthy attacks. In fact, the attacker may be satisfied with attack inputs that are difficult to detect, in the sense that the detector would need to collect more measurements to make a decision with a desired operating point of ROC. Although it is impractical to compute the exact values of these two probabilities for an arbitrary detector at every time k, we are able to apply the techniques in detection theory and information theory to obtain bounds for p_k^F and p_k^D. A classical example is the Chernoff-Stein Lemma [14]. This lemma characterizes the asymptotic exponent of p_k^F, while p_k^D can be arbitrary. Motivated by Chernoff-Stein Lemma, we propose the following notion of ϵ-stealthiness.

Definition* 3: *(ϵ-stealthiness) Let $\epsilon > 0$. An attack $\tilde{u}_1^\infty$ is ϵ-stealthy if, given any $0 < \delta < 1$, 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 $k \to \infty$. Namely, for any detector that satisfies $0 < 1 - p_k^D \leq \delta$ for all times k, we have

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

□

In fact, using Chernoff-Stein Lemma, we can provide a sufficient condition for an attack to be ϵ-stealthy.

Lemma* 1: *(Sufficient condition for ϵ-stealthiness) Suppose that an attack $\tilde{u}_1^\infty$ is such that the random sequence $\tilde{y}_1^\infty$ is ergodic and satisfies

$$\lim_{k\to\infty} \frac{1}{k} D\big(\tilde{y}_1^k\big\|y_1^k\big) \leq \epsilon. \tag{6}$$

Then the attack $\tilde{u}_1^\infty$ is ϵ-stealthy.

Proof: We apply the Chernoff-Stein Lemma for ergodic measurements (see, e.g. [19]). For such an attack $\tilde{u}_1^\infty$, given $0 < 1 - p_k^D \leq \delta$ where $0 < \delta < 1$, the best achievable exponent of p_k^F is given by $\lim_{k\to\infty} \frac{1}{k} D\big(\tilde{y}_1^k\big\|y_1^k\big)$. For any detector, we obtain

$$\limsup_{k\to\infty} -\frac{1}{k} \log p_k^F \leq \lim_{k\to\infty} \frac{1}{k} D\big(\tilde{y}_1^k\big\|y_1^k\big) \leq \epsilon.$$

By Definition 3, the attack is ϵ-stealthy. ■

Clearly, the measurement sequence $\tilde{y}_1^\infty$ to be ergodic is a limiting assumption on the attacker. Characterizing the level of stealthiness of a general non-ergodic attack seems to be hard because the Chernoff-Stein Lemma is not applicable in this case. Instead, we provide a necessary condition for an attack to be ϵ-stealthy regardless of the ergodicity of the measurement sequence.

Lemma* 2: *(Necessary condition for ϵ-stealthiness) If an attack $\tilde{u}_1^\infty$ is ϵ-stealthy, then

$$\limsup_{k\to\infty} \frac{1}{k} D\big(\tilde{y}_1^k\big\|y_1^k\big) \leq \epsilon. \tag{7}$$

The proof of Lemma 2 is postponed to the Appendix.

***Remark* 2:** Lemma 2 only provides a necessary condition of ϵ-stealthiness. If an attack satisfies (7), then it may not be ϵ-stealthy in general. □

We conclude this section with a method to compute the KLD between the sequences $\tilde{y}_1^k$ and y_1^k. For observed-based controllers, note that z_k and $\tilde{z}_k$ are functions of y_1^k and $\tilde{y}_1^k$, respectively. Using the invariance properties of KLD [17], we have

$$D\big(\tilde{y}_1^k \big\| y_1^k\big) = D\big(\tilde{z}_1^k \big\| z_1^k\big),$$

for every $k > 0$. Recall that z_1^∞ is an i.i.d. Gaussian random sequence with $z_k \sim \mathcal{N}(0, \sigma_z^2)$. From (4) we obtain

$$\frac{1}{k} D\big(\tilde{z}_1^k \big\| z_1^k\big) = -\frac{1}{k} h\big(\tilde{z}_1^k\big) + \frac{1}{2}\log(2\pi\sigma_z^2) + \frac{1}{k}\sum_{n=1}^{k} \frac{\mathbb{E}\big[\tilde{z}_n^2\big]}{2\sigma_z^2}, \tag{8}$$

where $h\big(\tilde{z}_1^k\big) = \int_{-\infty}^{\infty} -f_{\tilde{z}_1^k}(\xi_1^k) \log f_{\tilde{z}_1^k}(\xi_1^k) d\xi_1^k$ is the differential entropy [14] of z_1^k.

IV. Main Results

We are interested in the maximal performance degradation $\tilde{P}$ that an ϵ-stealthy attack may induce. We present such a fundamental limit in two parts: 1) the converse statement that gives an upper bound for $\tilde{P}$ as induced by the attacker, and 2) the achievability result that provides an attack that achieves the upper bound of the converse result.

Theorem* 1: *(Converse) Suppose that $\mathcal{I}_1^\infty$ satisfies (A1)–(A3). For any ϵ-stealthy attack $\tilde{u}_1^\infty$ generated by $\mathcal{I}_1^\infty$, we have

$$\tilde{P} \leq \bar{\delta}(\epsilon) P + \frac{(\bar{\delta}(\epsilon) - 1)\sigma_v^2}{c^2} \tag{9}$$

where the function $\bar{\delta} : [0, \infty) \to [1, \infty)$ is such that

$$\bar{\delta}(D) = 2D + 1 + \log \bar{\delta}(D). \tag{10}$$

Proof: Observe that $\tilde{z}_k = \tilde{y}_k - c\hat{\tilde{x}}_k = c(x_k - \hat{\tilde{x}}_k) + v_k$, and $(x_k - \hat{\tilde{x}}_k)$ is independent of v_k. We have

$$\mathbb{E}[\tilde{z}_k^2] = c^2 \tilde{P}_k + \sigma_v^2. \tag{11}$$

Since σ_v^2 is a constant and $c^2 > 0$, we can represent $\tilde{P}$ in terms of $\mathbb{E}[\tilde{z}_k^2]$. From (8), we have

$$\begin{aligned}
&\frac{1}{2}\cdot\frac{1}{k}\sum_{n=1}^{k}\frac{\mathbb{E}[\tilde{z}_n^2]}{\sigma_z^2} \\
&= \frac{1}{k} D\big(\tilde{z}_1^k \big\| z_1^k\big) - \frac{1}{2}\log(2\pi\sigma_z^2) + \frac{1}{k} h\big(\tilde{z}_1^k\big) \\
&\leq \frac{1}{k} D\big(\tilde{z}_1^k \big\| z_1^k\big) - \frac{1}{2}\log(2\pi\sigma_z^2) + \frac{1}{k}\sum_{n=1}^{k} h(\tilde{z}_n) \qquad (12) \\
&\leq \frac{1}{k} D\big(\tilde{z}_1^k \big\| z_1^k\big) - \frac{1}{2}\log(2\pi\sigma_z^2) + \frac{1}{k}\sum_{n=1}^{k} \frac{1}{2}\log\left(2\pi e\mathbb{E}[\tilde{z}_n^2]\right) \qquad (13) \\
&= \frac{1}{k} D\big(\tilde{z}_1^k \big\| z_1^k\big) + \frac{1}{2} + \frac{1}{2}\log\left(\prod_{n=1}^{k}\frac{\mathbb{E}[\tilde{z}_n^2]}{\sigma_z^2}\right)^{\frac{1}{k}} \\
&\leq \frac{1}{k} D\big(\tilde{z}_1^k \big\| z_1^k\big) + \frac{1}{2} + \frac{1}{2}\log\left(\frac{1}{k}\sum_{n=1}^{k}\frac{\mathbb{E}[\tilde{z}_n^2]}{\sigma_z^2}\right), \qquad (14)
\end{aligned}$$

where the inequalities (12) is due to the subadditivity of differential entropy [14, Corollary 8.6.1], the inequality (13) is a consequence of the maximum entropy theorem [14, Theorem 8.6.5], and the inequality (14) follows from the Arithmetic Mean and Geometric Mean (AM-GM) inequality. Consider the following maximization problem

$$\max_{x\in\mathbb{R}} x \quad \text{subject to } \frac{1}{2}x - D - \frac{1}{2} \leq \frac{1}{2}\log x \tag{15}$$

where $D \geq 0$. Since a logarithm function is concave, the feasible region of x in (15) is a closed interval upper bounded by $\bar{\delta}(D)$ as defined in (10); see Fig. 1. Thus, the maximum in (15) is $\bar{\delta}(D)$. By (14) and the maximization problem (15), we obtain

$$\frac{1}{k}\sum_{n=1}^{k}\frac{\mathbb{E}[\tilde{z}_n^2]}{\sigma_z^2} \leq \bar{\delta}\Big(\frac{1}{k} D\big(\tilde{z}_1^k \big\| z_1^k\big)\Big) \tag{16}$$

Using (11) and (16) gives

$$\begin{aligned}
\tilde{P} &= \limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\tilde{P}_n = \limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\frac{\mathbb{E}[\tilde{z}_n^2] - \sigma_v^2}{c^2} \\
&= \limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\frac{\bar{\delta}\Big(\frac{1}{n} D\big(\tilde{z}_1^n \big\| z_1^n\big)\Big)\sigma_z^2 - \sigma_v^2}{c^2} \\
&\leq \limsup_{k\to\infty}\frac{\bar{\delta}\Big(\frac{1}{k} D\big(\tilde{z}_1^k \big\| z_1^k\big)\Big)\sigma_z^2 - \sigma_v^2}{c^2} \qquad (17) \\
&= \frac{\bar{\delta}\Big(\limsup_{k\to\infty}\frac{1}{k} D\big(\tilde{z}_1^k \big\| z_1^k\big)\Big)\sigma_z^2 - \sigma_v^2}{c^2} \qquad (18) \\
&\leq \frac{\bar{\delta}(\epsilon)\sigma_z^2 - \sigma_v^2}{c^2} \qquad (19)
\end{aligned}$$

where the inequality (17) can be obtained by the definition of limit superior, the equality (18) is due to the continuity and monotonicity of the function $\bar{\delta}$, and the inequality (19) follows from Lemma 2. Finally, plugging $\sigma_z^2 = c^2 P + \sigma_v^2$ into (19) yields the desired result. ■

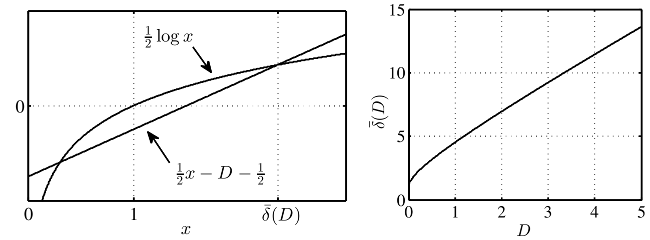

Fig. 1. Illustrations for the optimization problem (15) and the function $\bar{\delta} : [0, \infty)$ defined in (10). Notice that the function $\bar{\delta}$ is continuous and monotonically increasing.

Remark* 3: *(Effect of strictly stealthy attacks) Using (16), (11) and the fact that $\bar{\delta}(0) = 1$, if $D\big(\tilde{z}_1^k \big\| z_1^k\big) = 0$ is true for all $k > 0$, then $\mathbb{E}[\tilde{z}_k^2] = c^2\tilde{P}_k + \sigma_v^2 \leq c^2 P + \sigma_v^2$, which gives $\tilde{P}_k \leq P$. From Remark 1, we can conclude that an attack $\tilde{u}_1^\infty$ that is strictly stealthy can not degrade the MSE of the Kalman filter by any amount. □

We now present an ϵ-stealthy attack that achieves the upper bound in Theorem 1.

Theorem* 2: *(Achievability) The upper bound (9) in Theorem 1 is achievable by the attack $\tilde{u}_1^\infty$ generated by

$$\tilde{u}_k = u_k - (a - Kc)\zeta_{k-1} + \zeta_k, \tag{20}$$

where ζ_1^∞ is an i.i.d. sequence of random variables $\zeta_k \sim \mathcal{N}\big(0, \frac{\sigma_z^2}{c^2}(\bar{\delta}(\epsilon)-1)\big)$ independent of the system dynamics, and the initial condition is given by $\zeta_0 = 0$.

Proof: Notice that the attack in (20) can be generated by any information pattern satisfying (A1)–(A3). For ease of analysis, we assume that the attack $\tilde{u}_1^\infty$ is generated by an attacker with the information pattern $\mathcal{I}_1^\infty$ where $\mathcal{I}_k = \{u_1^k, \zeta_1^k, \tilde{y}_1^k\}$ for every $k > 0$.

We first show that the upper bound (9) is achieved by the attack. Moreover, the attacker implements the Kalman filter $\hat{x}_{k+1}^A = a\hat{x}_k^A + Kz_k^A + \tilde{u}_k$ with the initial condition $\hat{x}_1^A = 0$ where $z_k^A = \tilde{y}_k - c\hat{x}_k^A$. Therefore, $\hat{x}_{k+1}^A$ is the MMSE estimate of the state with the MSE $\mathbb{E}[(\hat{x}_{k+1}^A - x_{k+1})^2] = P$ when $\mathcal{I}_k$ is given. Note that $\tilde{z}_k$ can be expressed as

$$\tilde{z}_k = \tilde{y}_k - c\hat{\hat{x}}_k = \tilde{y}_k - c\hat{x}_k^A + c(\hat{x}_k^A - \hat{\hat{x}}_k) = z_k^A - c\tilde{e}_k \quad (21)$$

where $\tilde{e}_k = \hat{\hat{x}}_k - \hat{x}_k^A$. In addition, the dynamics of $\tilde{e}_k$ are given by

$$\begin{aligned}\tilde{e}_{k+1} &= (a\hat{\hat{x}}_k + K\tilde{z}_k + u_k) - (a\hat{x}_k^A + Kz_k^A + \tilde{u}_k)\\ &= (a-Kc)\tilde{e}_k + (a-Kc)\zeta_{k-1} - \zeta_k \qquad (22)\end{aligned}$$

and the initial condition is $\tilde{e}_1 = 0$. By induction, the equation (22) implies that $\tilde{e}_{k+1} = -\zeta_k$ for every $k > 0$. Further, for every $k > 0$, $\tilde{P}_{k+1}$ can be expressed as

$$\begin{aligned}\tilde{P}_{k+1} &= \mathbb{E}\big[(\hat{\hat{x}}_{k+1} - \hat{x}_{k+1}^A + \hat{x}_{k+1}^A - x_{k+1})^2\big]\\ &= \mathbb{E}\big[(\hat{\hat{x}}_{k+1} - \hat{x}_{k+1}^A)^2\big] + \mathbb{E}\big[(\hat{x}_{k+1}^A - x_{k+1})^2\big]\\ &\quad + 2\mathbb{E}\big[(\hat{\hat{x}}_{k+1} - \hat{x}_{k+1}^A)(\hat{x}_{k+1}^A - x_{k+1})\big] \qquad (23)\\ &= \mathbb{E}\big[(\tilde{e}_{k+1})^2\big] + P\\ &= \frac{\sigma_z^2}{c^2}(\bar{\delta}(\epsilon)-1) + P\\ &= \bar{\delta}(\epsilon)P + \frac{(\bar{\delta}(\epsilon)-1)\sigma_v^2}{c^2}. \qquad (24)\end{aligned}$$

In (23), the fact $\mathbb{E}\big[(\hat{\hat{x}}_{k+1} - \hat{x}_{k+1}^A)(\hat{x}_{k+1}^A - x_{k+1})\big] = 0$ is due to the principle of orthogonality, i.e., all the random variables generated by $\mathcal{I}_k$ is independent of the estimation error $(\hat{x}_{k+1}^A - x_{k+1})$ of the MMSE estimate. Hence, the upper bound of $\tilde{P}$ in (9) is achieved by this attack.

Now we show that the attack $\tilde{u}_1^\infty$ is ϵ-stealthy. From (21) and (22), we obtain $\tilde{z}_k = z_k^A - c\zeta_{k-1}$. Since $\{z_k^A\}_{k=1}^\infty$ is an i.i.d. random sequence with $z_k^A \sim \mathcal{N}(0, \sigma_z^2)$, the random sequence $\tilde{z}_1^\infty$ is i.i.d. Gaussian with $\tilde{z}_k \sim \mathcal{N}(0, \bar{\delta}(\epsilon)\sigma_z^2)$. For every $k > 0$, we can calculate the KLD by

$$\begin{aligned}\frac{1}{k}\sum_{n=1}^{k} D\big(\tilde{y}_1^k \| y_1^k\big) &= \frac{1}{k}\sum_{n=1}^{k} D\big(\tilde{z}_1^k \| z_1^k\big)\\ &= \frac{1}{k}\sum_{n=1}^{k} -\frac{1}{2}\log\big(2\pi e\bar{\delta}(\epsilon)\sigma_z^2\big) + \frac{1}{2}\log(2\pi\sigma_z^2) + \frac{\bar{\delta}(\epsilon)\sigma_z^2}{2\sigma_z^2}\\ &= -\frac{1}{2} - \frac{1}{2}\log\bar{\delta}(\epsilon) + \frac{1}{2}\bar{\delta}(\epsilon)\\ &= \epsilon\end{aligned}$$

where the differential entropy of $\tilde{z}_1^k$ is given by $h(\tilde{z}_1^k) = \sum_{n=1}^{k} h(\tilde{z}_n) = \frac{k}{2}\log\big(2\pi e\bar{\delta}(\epsilon)\sigma_z^2\big)$ because $\tilde{z}_1^\infty$ is an i.i.d. Gaussian sequence. In this case, $\tilde{y}_1^\infty$ is ergodic. From Lemma 1, the attack $\tilde{u}_1^\infty$ is ϵ-stealthy.

Finally, notice that the attack (20) can be generated by any information pattern that satisfies (A1)–(A3). Therefore, the converse result in Theorem 1 is achievable. ■

Remark* 4: *(Properties of attack* (20)*) In the proof of Theorem 1, the inequalities (12), (13), and (14) hold with equalities if and only if $\tilde{z}_1^k$ is a sequence of independent, Gaussian with mean zero, random variables stationary in the second moment, respectively. Clearly, the attack (20) satisfies all these conditions. □

Remark* 5: *(Attacker information pattern) Intuitively, the more information about the state variables an attacker has, the larger performance degradation it can induce. However, Theorem 1 and Theorem 2 imply that the only critical piece of information for the attacker is the nominal control input u_1^∞, due to the causality assumption of the information pattern and the i.i.d. property of the innovation $\tilde{z}_1^\infty$ required to achieve the upper bound (9). Note that knowledge of the nominal control input may not be necessary for different attack models. For instance, in the case the control input is transmitted via an additive channel, the attacker may achieve the upper bound (9) exploiting the linearity of the system, and without knowing the nominal control input. □

V. Numerical Results

We present numerical results to illustrate the fundamental performance bounds derived in Section IV. The following numerical results are in terms the ratio $\tilde{P}/P$, which can be interpreted as the attacker's gain. If the ratio $\tilde{P}/P = 1$, then the attacker can induce no degradation of the MSE.

Theorems 1 and Theorem 2 have formalized the notion that the stealthiness of an attacker can be a traded-off with the performance degradation it can induce. To illustrate such a trade-off numerically, we plot in Figure 2 the ratio $\tilde{P}/P$ as a function of ϵ where the system parameters are given by $a = 2$, $c = 1$, $\sigma_w^2 = 0.5$, $\sigma_v^2 = 0.1$, and $\tilde{P}$ is induced by the optimal attack in (20). We can see that if the attacker induces a larger performance degradation of the Kalman filter, then the attacker is detectable in fewer time steps.

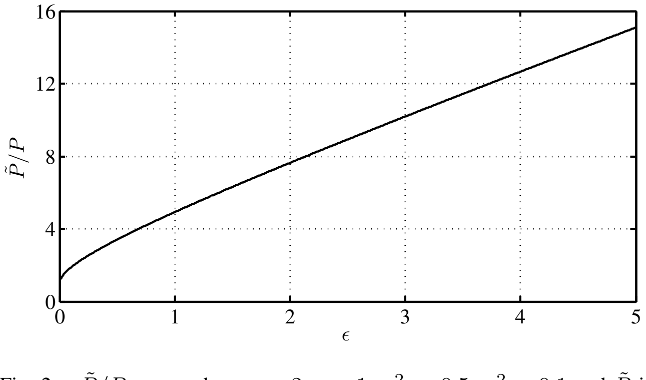

Fig. 2. $\tilde{P}/P$ vs. ϵ, where $a = 2$, $c = 1$, $\sigma_w^2 = 0.5$, $\sigma_v^2 = 0.1$ and $\tilde{P}$ is induced by the optimal ϵ-stealthy attack in (20).

Now we vary the quantity c^2/σ_v^2 and plot the ratio $\tilde{P}/P$. The quantity c^2/σ_v^2 can be viewed as the quality of the measurements taken by the sensor, i.e., the signal-to-noise ratio. In Fig. 3, we set $a = 2, \sigma_w^2 = 0.5$ and $\tilde{P}$ is induced by the optimal ϵ-stealthy attack in (20). Note that P is a function of c^2/σ_v^2 as well. We can see that a sensor taking measurements with a large value of c^2/σ_v^2 can lower the attacker's gain $\tilde{P}/P$. Consider the limiting condition $c^2/\sigma_v^2 \to 0^+$, i.e., the system becomes unobservable. In this case, the open loop unstable system is non-detectable and hence $P \to \infty$. In addition, we can take the limit of (9) and obtain $\tilde{P} \to \infty$ as $c^2/\sigma_v^2 \to 0^+$. Nevertheless, in Fig. 3, the attacker's gain $\tilde{P}/P$ remains bounded even if $\tilde{P}$ is unbounded as $c^2/\sigma_v^2 \to 0^+$.

On the other hand, we consider an open loop stable system ($|a| < 1$) with $a = 0.5$ and $\sigma_w^2 = 0.5$. The attacker's gain $\tilde{P}/P$ versus c^2/σ_v^2 for the system is presented in Fig. 4. We can see the similar effect if the quantity c^2/σ_v^2 is large. However, Fig. 4 shows that the attacker's gain for a stable system behaves differently from what we found in Fig. 3 as $c^2/\sigma_v^2 \to 0^+$. The stability assumption ($|a| < 1$) of the open loop system implies the boundedness of the MSE of the Kalman filter P for all $c^2/\sigma_v^2 \geq 0$. Taking the limit of (9), we find that $\tilde{P}$ goes to infinity as $c^2/\sigma_v^2 \to 0^+$. This explains the unboundedness of the attacker's gain $\tilde{P}/P$ as $c^2/\sigma_v^2 \to 0^+$.

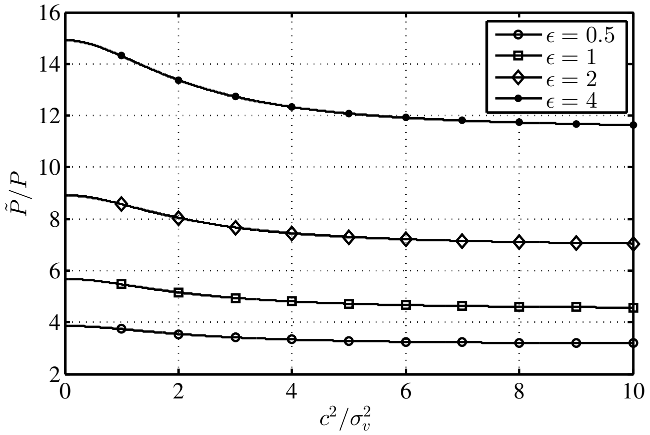

Fig. 3. $\tilde{P}/P$ vs. c^2/σ_v^2, where $a = 2$, $\sigma_w^2 = 0.5$ and $\tilde{P}$ is induced by the optimal ϵ-stealthy attack in (20).

VI. Conclusion

This work characterizes fundamental limitations and performance bounds for the security of stochastic control systems. The scenario is considered where the attacker knows the system parameters and noise statistics, and is able to hijack and replace the nominal control input. We propose a notion of ϵ-stealthiness to quantify the difficulty to detect an attack from measurements, and we characterize the largest degradation of the control performance induced by an ϵ-stealthy attack. Finally, our study reveals that an ϵ-stealthy attacker must know the nominal control input to cause the largest performance degradation of the control performance.

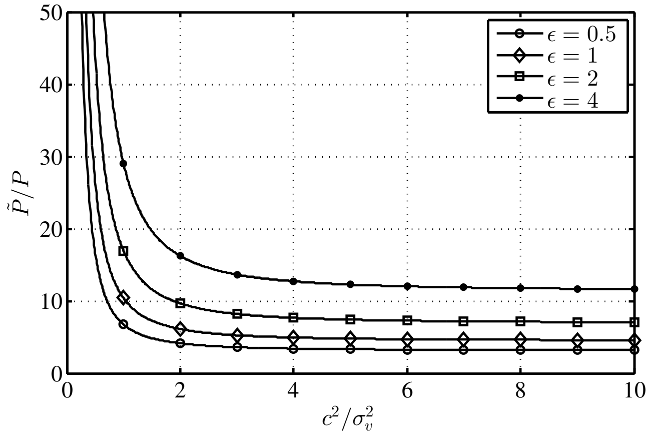

Fig. 4. $\tilde{P}/P$ vs. c^2/σ_v^2, where $a = 0.5$, $\sigma_w^2 = 0.5$ and $\tilde{P}$ is induced by the optimal ϵ-stealthy attack in (20).

Appendix
Proof of Lemma 2

We prove the lemma by contradiction. Assume that there exists an ϵ-stealthy attack $\tilde{u}_1^\infty$ such that $\limsup_{k\to\infty} \frac{1}{k} D\big(\tilde{y}_1^k \big\| y_1^k\big) > \epsilon$. Suppose that the detector employs the log-likelihood ratio test with a certain threshold λ_k at every time k, i.e.,

$$L_k(\eta_1^k) \underset{H_1}{\overset{H_0}{\lesseqgtr}} \lambda_k, \quad \text{where} \quad L_k(\eta_1^k) = \log \frac{f_{\tilde{y}_1^k}(\eta_1^k)}{f_{y_1^k}(\eta_1^k)}$$

is the log-likelihood ratio and $\eta_1^k = y_1^k$ (resp. $\eta_1^k = \tilde{y}_1^k$) if H_0 (resp. H_1) is true. We now use the technique due to Chernoff to bound p_k^F. Define the conditional cumulant generating function for the log-likelihood ratio to be $g_{k|0}(s) = \log \mathbb{E}\big[e^{sL_k}\big|H_0\big]$ and $g_{k|1}(s) = \log \mathbb{E}\big[e^{sL_k}\big|H_1\big]$. Note that $g_{k|0}(s) = g_{k|1}(s-1)$. Suppose that λ_k is chosen such that $0 < 1 - p_k^D \leq \delta$ for every $k > 0$. Then, for any $s_k > 0$, applying Chernorff's inequality yields

$$p_k^F = \mathbb{P}[L_k \geq \lambda_k | H_0] \leq e^{-s_k\lambda_k + g_{k|0}(s_k)},$$

and hence

$$\begin{aligned} -\log p_k^F &\geq s_k\lambda_k - g_{k|0}(s_k) \\ &= s_k\lambda_k - g_{k|1}(s_k - 1) \\ &= s_k\lambda_k - \log \mathbb{E}\big[e^{(s_k-1)L_k}\big|H_1\big] \\ &\geq s_k\lambda_k + \log \mathbb{E}\big[e^{-(s_k-1)L_k}\big|H_1\big] \qquad (25) \\ &\geq s_k\lambda_k + \mathbb{E}[-(s_k-1)L_k|H_1] \qquad (26) \\ &= D\big(\tilde{y}_1^k\big\|y_1^k\big) + s_k\Big(\lambda_k - D\big(\tilde{y}_1^k\big\|y_1^k\big)\Big), \qquad (27) \end{aligned}$$

where (25) and (26) follow from Jensen's inequality, and (27) is due to $\mathbb{E}[L_k|H_1] = D\big(\tilde{y}_1^k\big\|y_1^k\big)$. We choose $s_k > 0$ to be

$$s_k = \frac{1}{2}\left|\frac{D\big(\tilde{y}_1^k\big\|y_1^k\big) - k\epsilon}{D\big(\tilde{y}_1^k\big\|y_1^k\big) - \lambda_k}\right|. \qquad (28)$$

Using (27), (28) and the fact that $\limsup_{k\to\infty} \frac{1}{k} D\big(\tilde{y}_1^k\big\|y_1^k\big) > \epsilon$, we obtain

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

which contradicts to (5) because of the assumption of ϵ-stealthiness. Hence, the condition stated in (7) must be true.

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] H. S. Foroush and S. Martínez, “On multi-input controllable linear systems under unknown periodic dos jamming attacks.” in *SIAM Conf. on Control and its Applications*. SIAM, 2013, pp. 222–229.

[3] R. S. Smith, “A decoupled feedback structure for covertly appropriating networked control systems,” *Network*, vol. 6, p. 6, 2011.

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

[5] 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.

[6] 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, p. 13, 2011.

[7] 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.

[8] Y. Mo, R. Chabukswar, and B. Sinopoli, “Detecting integrity attacks on scada systems,” *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, 2014.

[9] G. Basile and G. Marro, *Controlled and Conditioned Invariants in Linear System Theory*. Prentice Hall, 1991.

[10] F. Pasqualetti, F. Dörfler, and F. Bullo, “Attack detection and identification in cyber-physical systems,” *IEEE Transactions on Automatic Control*, no. 11, 2013.

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

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

[13] C.-Z. Bai and V. Gupta, “On kalman filtering in the presence of a compromised sensor: Fundamental performance bounds,” in *American*, Portland, OR, June 2014, pp. 3029–3034.

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

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

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

[17] S. Kullback, *Information theory and statistics*. Courier Dover Publications, 1997.

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

[19] Y. Polyanskiy and Y. Wu, *Lecture notes on Information Theory*. MIT (6.441), UIUC (ECE 563), 2012–2013.