Stealthy Actuator Signal Attacks in Stochastic Control Systems: Performance and Limitations

Chongrong Fang, *Student Member, IEEE*, Yifei Qi, Jiming Chen, *Fellow, IEEE*, Rui Tan, *Senior Member, IEEE*, and Wei Xing Zheng, *Fellow, IEEE*

***Abstract*—In this technical note, the trade-off between the attack detectability and the performance degradation in stochastic cyber-physical systems is investigated. We consider a linear time-invariant system in which the attack detector performs a hypothesis test on the innovation of the Kalman filter to detect malicious tampering with the actuator signals. We adopt a notion of attack stealthiness to quantify the degree of stealth by limiting the maximum achievable exponents of both false alarm probability and detection probability below certain thresholds. And the conditions for any actuator attack to have a specific level of stealthiness are derived. Additionally, we characterize the upper bound of the performance degradation induced by attacks with a given extent of stealthiness that produce independent and identically distributed Gaussian innovations, and design the attack which achieves the stated upper bound for right-invertible systems. Finally, our results are illustrated via numerical examples.**

***Index Terms*—Cyber-physical system, Kalman filter, security.**

I. Introduction

Cyber-physical systems (CPSs) are systems with tight integration of the computational and physical components, which are currently widely used in modern society and attractive to attackers due to their significance. Both academia research [1, 2] and practical attack incidents such as the Stuxnet [3] and the Maroochy water breach [4] have demonstrated the feasibility of degrading the system performance by injecting malicious data into the communication channels. Thus, it is important to study the effects of data injection attack on control systems and develop countermeasures.

Different from contingencies and accidental malfunctions, the attacks aim to remain undetected while degrading the performance of the system. There are a series of research works [5–8] concentrating on the study of the well-crafted and stealthy attacks for certain detectors such as the χ^2 detector. In these works, the attackers properly manipulate sensor measurements or control commands based on the knowledge of the system and the detector. In addition, a more stealthy attack proposed in [9] strategically adapts to the time-varying detection threshold. As these attacks are designed for a certain detector, it falls short of characterizing the fundamental aspects of the attack, such as attack detectability. In practice, it is usually difficult for the attackers to obtain detailed information about the detectors. As a result, a notion of attack stealthiness that is independent of the details of the attack detector has received research attention recently.

In deterministic CPSs, to quantify the attack stealthiness with zero knowledge of detectors, Pasqualetti *et al.* [2] and Sundaram *et al.* [10] have shown that the modified sensor measurements will be treated as normal if and only if they excite the zero dynamics of the system. This zero-dynamics tampering strategy is independent of the attack detection algorithm. For stochastic control systems taking the process and measurement noises into consideration, Bai *et al.* introduce a notion of ϵ-stealthiness for false data injection (FDI) attacks on actuators [11, 12] and sensors [13]. These studies relate the attack stealthiness with the upper bound of the exponent of false alarm probability among arbitrary detectors. Kung *et al.* [14] study the difference of performance degradation induced by ϵ-stealthy attacks in the scalar and higher dimensional systems. Zhang *et al.* [15] consider the stealthiness in Linear Quadratic Gaussian (LQG) control systems and design stealthy FDI attacks over a finite time horizon. The work in [16] extends the ϵ-stealthiness to innovation-based linear attacks which are generated by manipulating sensor measurements and shows that the worst-case linear attack is zero-mean Gaussian distributed.

The Chernoff regime in information theories [17] demonstrates that the probabilities of false alarm and detection can converge exponentially fast at the same time when detectors perform sequential hypothesis tests. The notion of ϵ-stealthiness [11] in stochastic CPSs only considers the convergence rate of false alarm probability to be less than a given threshold. However, in most detectors that are widely used in CPSs, e.g., χ^2 detectors, the false alarm probability is usually fixed (i.e., its convergence rate is zero) while the convergence rate of detection probability could be large. It means that the attack may not be stealthy to such detectors, since the detection probability can converge to one very quickly while the false alarm probability remains constant. In addition, given the same level of ϵ-stealthiness, the attacks that result in faster convergence of detection probabilities are less stealthy (see Fig. 1). Therefore, we are motivated to consider a new notion of (ϵ, δ)-stealthiness for data injection attacks which provides a more general quantification of attack detectability in stochastic CPSs. As both the actuation and measurement communication channels are vulnerable in wireless [4] and wired [18] control systems, studying security issues in either channel contributes

This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0803501, the National Natural Science Foundation of China under Grant 61833015, and the NSW Cyber Security Network in Australia under Grant P00025091.

C. Fang, J. Chen, and Y. Qi are with the State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou, China (e-mail: chongrongfang.zju@gmail.com; cjm@zju.edu.cn; yifeiqi1127@gmail.com).

R. Tan is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore (e-mail: tanrui@ntu.edu.sg).

W. X. Zheng is with the School of Computing, Engineering and Mathematics, Western Sydney University, Sydney, NSW 2751, Australia (e-mail: w.zheng@westernsydney.edu.au).

to enhancing the security of CPSs. In this technical note, we focus on analyzing the trade-off between the performance degradation of the Kalman filter and the stealthiness level of attacks when the actuator signals of a linear time-invariant system are compromised. From our analysis and simulation results, compared with the ϵ-stealthiness, the proposed (ϵ, δ)-stealthiness presents a more fine-grained quantification of the attack stealthiness especially when the parameters ϵ and δ are within a certain range. The main contributions of this technical note are as follows:

(1) We use the Kullback-Leibler Divergence (KLD) [19] to characterize the attack stealthiness in stochastic CPSs which considers the convergence rates of false alarm probability and detection probability. Then, we derive the conditions for an attack to be (ϵ, δ)-stealthy.
(2) For the attacked innovations that are modeled as independent and identically distributed (i.i.d.) Gaussians, we derive the upper bound of the minimum mean-square estimation error (MMSE) of sensor measurements for (ϵ, δ)-stealthy attacks.
(3) We design the (ϵ, δ)-stealthy attacks that achieve the analytical maximum performance degradation in right-invertible systems. We conduct simulations to illustrate our results.

Notations: $\mathbb{R}^n$ is the n-dimensional Euclidean space. I_n means an $n \times n$ identity matrix. $\mathrm{tr}(\cdot)$ denotes the trace operation of a matrix. For any two vectors $\boldsymbol{x} = [x_1, \ldots, x_n]^T \in \mathbb{R}^n$ and $\boldsymbol{y} = [y_1, \ldots, y_n]^T \in \mathbb{R}^n$, $\boldsymbol{x} \leq \boldsymbol{y}$ (or $\boldsymbol{x} < \boldsymbol{y}$) means that $x_i \leq y_i$ (or $x_i < y_i$) for all $i = 1, \ldots, n$. And x_i^j is used to denote the sequence $\{x_n\}_{n=i}^j$. A Gaussian process with mean μ and covariance Σ is represented by $\mathcal{N}(\mu, \Sigma)$. The KLD between two random sequences r_1^k and s_1^k is defined by

$$D_{KL}(r_1^k \| s_1^k) = \int_{\xi_1^k \in \Phi} f_{r_1^k}(\xi_1^k) \log \frac{f_{r_1^k}(\xi_1^k)}{f_{s_1^k}(\xi_1^k)} d\xi_1^k, \quad (1)$$

where $\Phi = \{\xi_1^k | f_{r_1^k}(\xi_1^k) > 0\}$, and $f_{r_1^k}(\cdot)$ and $f_{s_1^k}(\cdot)$ are the joint probability density functions of r_1^k and s_1^k, respectively.

II. Problem formulation

A. System Model

Consider a discrete-time linear time-invariant (LTI) system:

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + w_k, \\ y_k &= Cx_k + v_k, \end{aligned} \quad (2)$$

where $x_k \in \mathbb{R}^n$ is the system state, $u_k \in \mathbb{R}^p$ is the actuator signal, $y_k \in \mathbb{R}^m$ is the sensor measurement, and A, B, C are known time-invariant matrices of appropriate dimensions. The $w_k \in \mathbb{R}^n$ and $v_k \in \mathbb{R}^m$ are uncorrelated zero-mean Gaussian random noises with covariance $Q \geq 0$ and $R \geq 0$, respectively. The initial state $x_1 \sim \mathcal{N}(0, P_1)$, where $P_1 \geq 0$, is independent of w_k and v_k for all $k \geq 1$. We assume that the pairs (A, B) and $(A, \sqrt{Q})$ are controllable and (A, C) is observable, and the system $\{A, B, C\}$ is right-invertible which is common in linear systems with feedback control [20].

Denote $y_1^k = \{y_1, \ldots, y_k\}$ as the measurements collected by the sensors from time 1 to time k. To estimate the system state, the Kalman filter is employed to perform the MMSE estimation of x_k from the historical sensor measurements y_1^{k-1}. Based on the observability and controllability conditions mentioned above, the Kalman filter converges exponentially to a steady-state [21]. In the following, it is assumed that the MMSE estimate $\hat{x}_k$ is calculated by a steady-state Kalman filter with the initial estimate $\hat{x}_1 = 0$, which is given as:

$$\begin{aligned} \hat{x}_{k+1} &= A\hat{x}_k + Kz_k + Bu_k, \\ K &= APC^T(CPC^T + R)^{-1}, \\ P &= APA^T - APC^T(CPC^T + R)^{-1}CPA^T + Q, \end{aligned}$$

where the innovation $z_k \triangleq y_k - C\hat{x}_k$ is an i.i.d. Gaussian process with mean zero and covariance $\Sigma_z = CPC^T + R$.

B. Attack Model

Considering the vulnerability of communication links and the ever-increasing attack capabilities, we employ an attack model in which the attackers can replace the actuator signals u_1^∞ with an arbitrary sequence $\tilde{u}_1^\infty$. The attack is designed based on the available system knowledge. Denote Γ_k as the set of obtainable system information of attackers at time k, and Γ_k should satisfy the following assumptions:

(A1) the system parameters $\{A, B, C, Q, R\} \in \Gamma_k$,
(A2) the actuator signal $u_k \in \Gamma_k$ at all time k,
(A3) Γ_k is non-decreasing (i.e., $\Gamma_k \subseteq \Gamma_{k+1}$) and Γ_k is independent of w_k^∞ and v_{k+1}^∞ for all k.

***Remark** 1:* Assumptions (A1) and (A2) are both reasonable since sophisticated attackers can obtain the system knowledge through first-principle modeling or system identification methods. The feasibility of compromising actuator signals has been shown in the literature and incidents such as [3, 4, 18]. The assumption (A3) stems from causality constraints.

Denote $\hat{\tilde{x}}_k$ and $\tilde{y}_k$ as the state estimate and sensor measurement under the attack $\tilde{u}_1^{k-1}$, respectively. As the system does not know $\tilde{u}_1^\infty$, the Kalman filter under the attack evolves as

$$\hat{\tilde{x}}_{k+1} = A\hat{\tilde{x}}_k + K\tilde{z}_k + Bu_k, \quad (3)$$

with corrupted innovation $\tilde{z}_k = \tilde{y}_k - C\hat{\tilde{x}}_k$. Notice that $\hat{\tilde{x}}_k$ is no longer the optimal MMSE estimate of the true state x_k now.

The accuracy of the estimation or prediction of system states and sensor measurements is important for applications such as state-based control, monitoring and so on. Hence, we assume that the attackers aim to degrade the system performance by increasing the covariance of the error between the predicted sensor data $\hat{\tilde{y}}_k$ and the true value y_k. Furthermore, we weight each element of the error vector $\hat{\tilde{y}}_k - y_k$ to normalize the relative attack impact among different elements of this error vector, and consider

$$J = \limsup_{k \to \infty} \frac{1}{k} \sum_{n=1}^{k} \mathbb{E}\left[\left(\hat{\tilde{y}}_n - y_n\right)^T \Sigma_z^{-1} \left(\hat{\tilde{y}}_n - y_n\right) \right]$$

as the performance metric [12]. This metric is the average normalized error covariance of sensor measurements over an infinite time horizon. In addition, J can also be formulated as

$$J = \limsup_{k \to \infty} \frac{1}{k} \sum_{n=1}^{k} \mathrm{tr}(\tilde{P}_k W) + \mathrm{tr}(\Sigma_z^{-1} R), \quad (4)$$

where $\tilde{P}_k = \mathbb{E}[(\hat{\tilde{x}}_k - x_k)(\hat{\tilde{x}}_k - x_k)^T]$ and $W = C^T\Sigma_z^{-1}C$.

C. Detector Description

The system detectors resort to analyzing the received sensor measurements to determine whether there is an attack in the system. We formulate the problem of detecting attacks as a sequential hypothesis testing problem. Specifically, at time k, the detector obtains the sensor data $y_1^k = \{y_1, \ldots, y_k\}$ and performs the following binary hypothesis test:

H_0 : No attack exists (the estimator receives y_1^k),
H_1 : Attack exists (the estimator receives $\tilde{y}_1^k$).

Note that we do not impose any restriction on the detection algorithm. For a given detector, we use detection probability p_k^D and false alarm probability p_k^F at time k to measure its detection performance, where p_k^D and p_k^F are given by

$$p_k^D = \mathrm{P}_k(H_1|H_1), \quad p_k^F = \mathrm{P}_k(H_1|H_0).$$

As the sensors collect data continuously and the detector performs a hypothesis test on all historical data, the convergence of p_k^F and p_k^D is possible. Given an attack and a detector, we denote E_1 and E_2 as the exponential convergence rates of p_k^F and p_k^D as $k \to \infty$, respectively. Specifically, we have

$$E_1 = \limsup_{k\to\infty} -\frac{1}{k}\log p_k^F, \quad E_2 = \limsup_{k\to\infty} -\frac{1}{k}\log(1-p_k^D).$$

Note that $E_1 \geq 0$ and $E_2 \geq 0$ for all detectors and the equality may hold, for example, $E_1 = 0$ in the constant false alarm probability detectors. Furthermore, we define the optimal trade-off between E_1 and E_2 by $E_2^*(E_1)$ as follows:

$$\begin{aligned} E_2^*(E_1) = \sup\{&E_2 : \exists k^0, \forall k \geq k^0, \\ &\exists \text{ detector } s.t.\ p_k^F < 2^{-kE_1},\ p_k^D > 1 - 2^{-kE_2}\}, \end{aligned}$$

where $E_2^*(E_1)$ is the maximum convergence rate of p_k^D among the detectors whose exponential convergence rate of p_k^F is E_1 as $k \to \infty$, and $E_2^*(E_1)$ is decreasing with respect to E_1.

D. Motivation and Problem Statement

Without the knowledge of the aforementioned detectors, e.g., the type or threshold of the detector, it is difficult for the attackers to guarantee that the crafted attacks are undetectable. The design of stealthy attacks will follow a random guessing approach. If the designed attack sequence does not trigger the alarm condition of the detector used, it can fortunately bypass the detector. As a result, with the observation that the attackers in the real world may wish to have long enough stealthy time to achieve attack objectives instead of being noticed all the time, it is reasonable to characterize the stealthiness of attacks against any detector by the increasing rate of p_k^D or the decay rate of p_k^F over time. Faster rates mean lower stealthiness. The recently proposed ϵ-stealthiness, as Definition 1 shows, connects the attack stealthiness with the convergence rate of p_k^F based on detection and information theories.

***Definition** 1 (ϵ-stealthiness [11]):* Let $\epsilon > 0$ and $0 < \theta < 1$. The attack $\tilde{u}_1^\infty$ is ϵ-stealthy if for any detector that operates with $0 < 1 - p_k^D \leq \theta$ at all time k, the following holds:

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

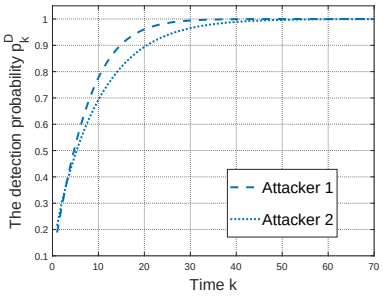

Fig. 1. The detection probabilities for two attacks, where the normal Kalman filter innovation: $z_k \sim \mathcal{N}_0(0, 0.621)$, the innovation of attacker 1: $\tilde{z}_k^1 \sim \mathcal{N}_1(0.6, 0.621)$ and the innovation of attacker 2: $\tilde{z}_k^2 \sim \mathcal{N}_2(0, 1.5483)$. Both attacks are ϵ-stealthy according to the mean and variance of $\tilde{z}_k^1$ and $\tilde{z}_k^2$ [11], where $\epsilon = 0.2816$. In addition, a detector in which $p_k^F = 0.05$ is adopted.

However, as it is stated in [17], the detector can make decisions on E_1 and E_2 simultaneously where $E_2 \leq E_2^*(E_1)$. Actually, the detectors that are widely used in real CPSs such as the χ^2 detector and CUmulative SUM (CUSUM) detector, typically make the false alarm probability p_k^F constant over time, i.e., $E_1 = 0$. This engenders the possibility that the detection probability p_k^D converges to one at the maximum exponential convergence rate $E_2^*(E_1 = 0)$. In such case, the attack may not be stealthy, unless $E_2^*(E_1 = 0)$ is less than an acceptable value.

In addition, the ϵ-stealthiness may not be sufficient in characterizing the extent of attack stealthiness. As Fig. 1 shows, with the false alarm probability p_k^F fixed, the probabilities of detecting the two attackers converge to one at different rates. Since the detector works based on the Neyman-Pearson Lemma [22] (i.e., the most powerful likelihood-ratio test with constant false alarm probability), the detection probability p_k^D at each time is the best achievable value. Clearly, attacker 2 is more stealthy than attacker 1, although they are of the same stealthiness level according to the definition of ϵ-stealthiness.

Thus, it motivates us to consider a more comprehensive notion of (ϵ, δ)-stealthiness capturing both the convergence rates E_1 and E_2. The (ϵ, δ)-stealthiness will be formally defined in Section III. In addition, we assume that the objective of the attacker is to maximize the performance metric J while maintaining a specific level of stealthiness by replacing the nominal actuator signals with a malicious attack sequence $\tilde{u}_1^\infty$. Formally, the attacker aims to solve the following problem:

$$\mathbf{P_0}: \quad \max\ J, \quad s.t.\ \text{the attack } \tilde{u}_1^\infty \text{ is } (\epsilon, \delta)\text{-stealthy}.$$

In this technical note, we will also study the performance degradation caused by the defined stealthy attacks that generate i.i.d. Gaussian innovations and find the attack $\tilde{u}_1^\infty$ that achieves the largest performance degradation in right-invertible systems.

III. Definition and Conditions for Stealthy Attack

In this section, we present the definition of attack stealthiness in stochastic CPSs based on the convergence rates of false alarm probability and detection probability. We also derive the conditions for an attacker to achieve a particular extent of stealthiness.

A. Attack Stealthiness

For the same attack, different detectors have distinct detection performance when performing hypothesis testing on sensor data. In other words, the convergence rates (E_1, E_2) among detectors may be different. Intuitively, an attacker is strictly stealthy if no better detection performance than the random guessing can be found among all detectors. Similarly, for a given attack, if the convergence rates E_1 and E_2 are upper-bounded with any detector, the attack possesses a certain degree of stealthiness. We formalize this understanding as Definition 2.

***Definition** 2 (Stealthy attacks):* Considering the system (2) and the detector mentioned in Section II-C, the attack $\tilde{u}_1^\infty$ is

(1) strictly stealthy if no detector has the detection performance that $p_k^F < p_k^D$ for any $k > 0$.

(2) (ϵ, δ)-stealthy with $\epsilon > 0$ and $\delta > 0$, if all detectors satisfy the following two conditions simultaneously:

(i) For $0 < \theta_1 < 1$ and any detector that satisfies $0 < 1 - p_k^D \leq \theta_1$ for all time k, the false alarm probability p_k^F converges to zero exponentially fast with rate smaller than ϵ as $k \to \infty$, namely,

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

(ii) For $0 < \theta_2 < 1$ and any detector that satisfies $0 < p_k^F \leq \theta_2$ for all time k, the detection probability p_k^D converges to one exponentially fast with rate smaller than δ as $k \to \infty$, namely,

$$\limsup_{k\to\infty} -\frac{1}{k}\log(1 - p_k^D) \leq \delta.$$

In summary, Definition 2 means that for an attack to be (ϵ, δ)-stealthy, both the maximum exponential convergence rates of p_k^F and p_k^D of the attack among all possible detectors need to be constrained below ϵ and δ, respectively.

B. Conditions for Stealthy Attack

According to Definition 2, we give the following theorem.

***Theorem** 1:* For the system (2) with a detector making hypothesis tests in Section II-C, the attack sequence $\tilde{u}_1^\infty$ is

(1) strictly stealthy, if $D_{KL}(\tilde{y}_1^k \| y_1^k) = D_{KL}(y_1^k \| \tilde{y}_1^k) = 0$.

(2) (ϵ, δ)-stealthy for $\epsilon > 0$ and $\delta > 0$, if and only if

$$\limsup_{k\to\infty} \frac{1}{k} D_{KL}(\tilde{y}_1^k \| y_1^k) \leq \epsilon, \tag{5}$$

$$\limsup_{k\to\infty} \frac{1}{k} D_{KL}(y_1^k \| \tilde{y}_1^k) \leq \delta. \tag{6}$$

Proof: For the first statement, it is straightforward to obtain the proof from the Neyman-Pearson Lemma.

For the second statement, from Lemmas 1 and 2 in [11], (5) is the sufficient and necessary condition for an ϵ-stealthy attack which meets Definition 1. Since part (i) of Definition 2 is consistent with Definition 1, (5) is part of the sufficient and necessary conditions for (ϵ, δ)-stealthiness. Then, by changing the original hypothesis to $H_0^* = H_1$ and $H_1^* = H_0$ where

H_0^* : No attack exists (the estimator receives $\tilde{y}_1^k$),

H_1^* : Attack exists (the estimator receives y_1^k),

the false alarm probability and detection probability of the current hypothesis testing problem are $p_k^{F^*} = P_k(H_1^*|H_0^*) = 1 - p_k^D$ and $p_k^{D^*} = P_k(H_1^*|H_1^*) = 1 - p_k^F$, respectively. Let $\epsilon^* = \delta$. For an attack to be ϵ^*-stealthy in such detectors, we must have $\limsup_{k\to\infty} -\frac{1}{k}\log(1 - p_k^D) \leq \delta$ when $0 < p_k^F \leq \theta_2 < 1$ holds at any time k, which coincides with part (ii) of Definition 2. Similar to (5), (6) is the sufficient and necessary condition for an attack to satisfy part (ii) of Definition 2. In summary, by combining (5) and (6), the theorem follows. ■

***Remark** 2:* Theorem 1 gives sufficient and necessary conditions that can be used to check if the attack $\tilde{u}_1^\infty$ is strictly stealthy or (ϵ, δ)-stealthy. Note that the criteria in Theorem 1 can be applied to any attack sequence, i.e., regardless of the distribution of $\tilde{u}_1^k$ or $\tilde{y}_1^k$. Through Theorem 1, one can find that there is no attacker being $(0, \delta)$-stealthy or $(\epsilon, 0)$-stealthy with $\epsilon > 0$ and $\delta > 0$, since, for example, $D_{KL}(y_1^k \| \tilde{y}_1^k) = 0$ must hold if $D_{KL}(\tilde{y}_1^k \| y_1^k) = 0$. Thus, the strictly stealthy attack here is equivalent to the strictly stealthy attack in [11]. In addition, as $\tilde{z}_k$ and z_k are invertible functions of $\tilde{y}_k$ and y_k respectively, we have $D_{KL}(\tilde{y}_1^k || y_1^k) = D_{KL}(\tilde{z}_1^k || z_1^k)$ and $D_{KL}(y_1^k || \tilde{y}_1^k) = D_{KL}(z_1^k || \tilde{z}_1^k)$ for every $k > 0$ due to the invariance property of the KLD [19].

IV. Performance degradation under attacks

It is important to investigate the performance degradation under (ϵ, δ)-stealthy attacks. In the rest of this technical note, we consider that the innovation sequence $\tilde{z}_1^k$ induced by the (ϵ, δ)-stealthy attack $\tilde{u}_1^{k-1}$ is an i.i.d. Gaussian process, i.e., $\tilde{z}_k \sim \mathcal{N}(\mu, \Sigma_{\tilde{z}})$. In the following, we first present the trade-off between D^ϵ and D^δ, which is critical to obtaining the largest performance degradation. Then, the upper bound of the performance degradation in the presence of stealthy attacks is derived. Further, the (ϵ, δ)-stealthy attack that achieves the obtained upper bound is designed for right-invertible systems.

A. Trade-off between D^ϵ and D^δ

We define the following terms:

$$\tilde{\Xi}_k = \mathbb{E}[\tilde{z}_k \tilde{z}_k^T] = \mu\mu^T + \Sigma_{\tilde{z}},\ \boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^T,$$

$$D^\epsilon = \frac{1}{2}\sum_{i=1}^m \lambda_i - 1 - \log\lambda_i,\ D^\delta = \frac{1}{2}\sum_{i=1}^m \frac{1}{\lambda_i} - 1 - \log\frac{1}{\lambda_i},$$

where $\boldsymbol{\lambda}$ is the vector of the eigenvalues of the matrix $\tilde{\Xi}_k \Sigma_z^{-1}$. Note that $\tilde{\Xi}_k \geq \Sigma_{\tilde{z}} > 0$ since $\mu\mu^T$ is positive semi-definite. The following lemma is given as a preliminary.

***Lemma** 1 ([13], Lemma 10):* For any $x \geq 0$, we define

$$\bar{\delta}(x) = 2x + 1 + \log\bar{\delta}(x) \text{ where } \bar{\delta} : [0, \infty) \to [1, \infty),$$

$$\underline{\delta}(x) = 2x + 1 + \log\underline{\delta}(x) \text{ where } \underline{\delta} : [0, \infty) \to (0, 1].$$

Then, $\bar{\delta}(x)$ and $\underline{\delta}(x)$ are increasing concave and decreasing convex functions, respectively, when $x \geq 0$.

When (ϵ, δ)-stealthy attacks result in i.i.d. Gaussian innovations, D^ϵ and D^δ respectively represent the eigenvalue form of (5) and (6) in Theorem 1 after certain derivation, which will be described in detail in Section IV-B. In this subsection, the trade-off between D^ϵ and D^δ is provided in Lemma 2, which lays the basis for deriving the largest performance degradation.

***Lemma** 2:* Considering that the detectors mentioned in Section II-C perform hypothesis tests between the normal innovation $z_k \sim \mathcal{N}(0,\Sigma_z)$ and the corrupted innovation $\tilde{z}_k \sim \mathcal{N}(\mu,\Sigma_{\tilde{z}})$ induced by (ϵ,δ)-stealthy attacks where $\epsilon>0$ and $\delta>0$. If $D^{\epsilon}=\epsilon$, the feasible range of $D^{\delta}(\epsilon)$ is given by

$$D^{\delta}_{\min}(\epsilon) \le D^{\delta}(\epsilon) \le D^{\delta}_{\max}(\epsilon),$$
$$D^{\delta}_{\min}(\epsilon) = \epsilon + \frac{1}{2}\left[\frac{1}{\bar{\delta}(\epsilon)} - \bar{\delta}(\epsilon) + 2\log\bar{\delta}(\epsilon)\right],$$
$$D^{\delta}_{\max}(\epsilon) = \epsilon + \frac{m}{2}\left[\frac{1}{\bar{\delta}(\frac{\epsilon}{m})} - \bar{\delta}\left(\frac{\epsilon}{m}\right) + 2\log\bar{\delta}\left(\frac{\epsilon}{m}\right)\right].$$

Proof: Our goal is to get the maximum and minimum of D^{δ} subject to the constraint $D^{\epsilon}=\epsilon$, which can be formulated as Problem $\mathbf{P_1}$ based on the above formulas of D^{ϵ} and D^{δ}:

$$\mathbf{P_1}: \max_{\boldsymbol{\lambda}} (\min_{\boldsymbol{\lambda}}) \ D^{\delta}-D^{\epsilon} = \frac{1}{2}\sum_{i=1}^{m}\left(\frac{1}{\lambda_i}-\lambda_i+2\log\lambda_i\right),$$
$$s.t. \ \frac{1}{2}\sum_{i=1}^{m}(\lambda_i-1-\log\lambda_i)=\epsilon.$$

To achieve the goal, we transform Problem $\mathbf{P_1}$ by letting $\lambda_i=\bar{\delta}(\epsilon_i)$ and $s_i(\epsilon_i)=\frac{1}{2}\left(\frac{1}{\bar{\delta}(\epsilon_i)}-\bar{\delta}(\epsilon_i)+2\log\bar{\delta}(\epsilon_i)\right)$ for $i=1,2,\dots,m$, and then calculate the maximum and minimum, respectively. From Lemma 1, we have $\bar{\delta}(x)\in(1,\infty)$, $\bar{\delta}'(x)=2\bar{\delta}(x)/(\bar{\delta}(x)-1)$ and $\bar{\delta}''(x)=-4\bar{\delta}(x)/(\bar{\delta}(x)-1)^3$ for $x>0$. Then, for $\epsilon_i>0$, the following inequality holds:

$$\begin{aligned} s_i''(\epsilon_i) &= \frac{1}{2}\left[\frac{2-2\bar{\delta}(\epsilon_i)}{\bar{\delta}^2(\epsilon_i)}(\bar{\delta}'(\epsilon_i))^2-\bar{\delta}''(\epsilon_i)(1-\frac{1}{\bar{\delta}(\epsilon_i)})^2\right] \\ &= \frac{2}{(1-\bar{\delta}(\epsilon_i))\bar{\delta}(\epsilon_i)} < 0. \end{aligned}$$

Without loss of generality, assuming that $\phi_1=0$ and $\phi_2\ge 0$, we have that $\forall t\in[0,1]$, $s_i(t\phi_2+(1-t)\phi_1)\ge t\cdot s_i(\phi_2)$ since $s_i(0)=0$ and $s_i''(\epsilon_i)<0$ on $\epsilon_i>0$. In summary, $s_i(\epsilon_i)$ is concave on $\epsilon_i\ge 0$. To get $D^{\delta}_{\max}(\epsilon)$, we rewrite $\mathbf{P_1}$ as

$$\mathbf{P_2}: \max \sum_{i=1}^{m} s_i(\epsilon_i), \quad s.t. \ \sum_{i=1}^{m}\epsilon_i=\epsilon, \ \epsilon_i\ge 0.$$

Based on Jensen's inequality, $\mathbf{P_2}$ is solved and $D^{\delta}_{\max}(\epsilon)$ is obtained when $\lambda_i=\bar{\delta}(\frac{\epsilon}{m})$, $\epsilon_i=\frac{\epsilon}{m}$ for $i=1,\dots,m$.

To obtain $D^{\delta}_{\min}(\epsilon)$, it is equivalent to solving the problem

$$\mathbf{P_3}: \max \sum_{i=1}^{m} -s_i(\epsilon_i), \quad s.t. \ \sum_{i=1}^{m}\epsilon_i=\epsilon, \ \epsilon_i\ge 0,$$

where $-s_i(\epsilon_i)$ is convex for $\epsilon_i\ge 0$ and the set of the constraint is convex. Based on the maximum principle in Convex Analysis [23], the optimal solution of $\mathbf{P_3}$ must exist on the boundary. Without loss of generality, let the boundary of the constraint set be $\Omega=\left\{\sum_{i=1}^{m-1}\epsilon_i=\epsilon, \epsilon_m=0\right\}$, which is convex and can be interpreted as a line segment in an m-dimensional space. Consequently, $\mathbf{P_3}$ is still convex on Ω. Reusing the maximum principle, the current constraint Ω is reduced to a set of points, i.e., $\{\epsilon_i=\epsilon, \epsilon_j=0, j\ne i, \ j=1,\dots,m \mid i=1,\dots,m\}$, in which each point leads to $D^{\delta}_{\min}(\epsilon)$. ∎

***Corollary** 1:* From Lemma 2, the following statements hold: (i) given $\epsilon>0$ and $\delta>0$, if $\delta\le D^{\delta}_{\min}(\epsilon)$, then $D^{\epsilon}_{\max}(\delta)\le\epsilon$ follows, (ii) there exists an (ϵ,δ)-stealthy attack achieving both the equality of (5) and (6) if $\delta\in\left[D^{\delta}_{\min}(\epsilon), D^{\delta}_{\max}(\epsilon)\right]$.

Proof: The statement (i) follows by fixing $D^{\delta}=\delta$ and repeating similar procedures in Lemma 2. As D^{ϵ} and D^{δ} are respectively the eigenvalue form of (5) and (6) when $\tilde{z}_k$ is i.i.d. Gaussian, the statement (ii) is derived directly from Lemma 2 and Theorem 1. ∎

***Remark** 3:* Lemma 2 and Corollary 1 indicate that the (ϵ,δ)-stealthiness presents a more fine-grained characterization of attack stealthiness than the ϵ-stealthiness in [11]. Specifically, fixing the maximum convergence rate of p_k^F to ϵ (i.e., $D^{\epsilon}=\epsilon$), the feasible maximum convergence rate of p_k^D belongs to the interval $\left[D^{\delta}_{\min}(\epsilon), D^{\delta}_{\max}(\epsilon)\right]$ when $\tilde{z}_k$ is i.i.d. Gaussian. The attacks with the same ϵ and different $\delta\in\left[D^{\delta}_{\min}(\epsilon), D^{\delta}_{\max}(\epsilon)\right]$ have distinct (ϵ,δ)-stealthiness levels, while they are ϵ-stealthy.

B. Performance degradation under (ϵ,δ)-stealthy attacks

To analyze the performance degradation induced by (ϵ,δ)-stealthy attacks, firstly we have $\tilde{\Xi}_k=C\tilde{P}_kC^T+R$ based on the observation of $\tilde{z}_k=C(x_k-\hat{\tilde{x}}_k)+v_k$ and Assumption (A3) in Section II-B. Then, based on (4), the performance metric is given by

$$\begin{aligned} J &\overset{(a)}{=} \limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\mathrm{tr}(C\tilde{P}_kC^T\Sigma_z^{-1})+\mathrm{tr}(R\Sigma_z^{-1}) \\ &\overset{(b)}{=} \limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\mathrm{tr}\left(\tilde{\Xi}_k\Sigma_z^{-1}\right) \overset{(c)}{=} \mathrm{tr}(\tilde{\Xi}_k\Sigma_z^{-1}) = \sum_{i=1}^{m}\lambda_i, \end{aligned}$$

where (a) holds since the trace operator is invariant under the cyclic permutations, and (b) and (c) follow due to the equation $\tilde{\Xi}_k=C\tilde{P}_kC^T+R$ and the i.i.d. property of $\tilde{z}_k$, respectively.

For an attack to be (ϵ,δ)-stealthy, according to Theorem 1 and Remark 2, the following inequalities are obtained from (5) and (6) respectively:

$$\limsup_{k\to\infty}\frac{1}{k}D_{KL}(\tilde{z}_1^k\|z_1^k)=D_{KL}(\tilde{z}_k\|z_k)\le\epsilon, \tag{7}$$
$$\limsup_{k\to\infty}\frac{1}{k}D_{KL}(z_1^k\|\tilde{z}_1^k)=D_{KL}(z_k\|\tilde{z}_k)\le\delta. \tag{8}$$

Taking (8) as an example, we expand the term $D_{KL}(z_k\|\tilde{z}_k)$ according to the definition of the KLD and then have

$$\begin{aligned} &\frac{1}{m}\mathrm{tr}(\tilde{\Xi}_k^{-1}\Sigma_z) \\ &= \frac{2}{m}D_{KL}(z_k\|\tilde{z}_k)+1+\frac{1}{m}\log|\tilde{\Xi}_k^{-1}\Sigma_z| \\ &- \frac{1}{m}\mathrm{tr}\left((\Sigma_{\tilde{z}}^{-1}-\tilde{\Xi}_k^{-1})\Sigma_z\right)-\frac{1}{m}D_{KL}\left(\mathcal{N}(0,\tilde{\Xi}_k)\|\mathcal{N}(0,\Sigma_{\tilde{z}})\right) \\ &\le \frac{2}{m}D_{KL}(z_k\|\tilde{z}_k)+1+\frac{1}{m}\log|\tilde{\Xi}_k^{-1}\Sigma_z|, \end{aligned}$$

where the inequality holds since $\tilde{\Xi}_k\ge\Sigma_{\tilde{z}}>0$ and every KLD is non-negative. To achieve the equality, we need $\mu=0$. Similarly, $D_{KL}(\tilde{z}_k\|z_k)$ satisfies the following:

$$\begin{aligned} &\frac{1}{m}\mathrm{tr}(\tilde{\Xi}_k\Sigma_z^{-1}) \\ &= \frac{2}{m}D_{KL}(\tilde{z}_k\|z_k)+1+\frac{1}{m}\log|\tilde{\Xi}_k\Sigma_z^{-1}|-\frac{1}{m}\log|\tilde{\Xi}_k\Sigma_{\tilde{z}}^{-1}| \\ &\le \frac{2}{m}D_{KL}(\tilde{z}_k\|z_k)+1+\frac{1}{m}\log|\tilde{\Xi}_k\Sigma_z^{-1}|. \end{aligned}$$

Considering a strictly stealthy attack and a normal case, the resulting innovation is $\tilde{z}_k \sim \mathcal{N}(0, \Sigma_z)$ by Theorem 1 and $z_k \sim \mathcal{N}(0, \Sigma_z)$, respectively. In both cases, we have $J = m$ since each $\lambda_i = 1$ now. As the adversary aims to increase J, we consider $\tilde{\Xi}_k \geq \Sigma_z$ for attackers, which leads to $\lambda_i \geq 1$ for $i = 1, \ldots, m$. As a result, we find that obtaining the upper bound of the performance degradation under (ϵ, δ)-stealthy attacks is equivalent to solving the following Problem $\mathbf{P_4}$, where (d) and (e) are respectively obtained by translating (7) and (8) with the eigenvalues of $\tilde{\Xi}_k \Sigma_z^{-1}$:

$$\begin{aligned} \mathbf{P_4}: \ \max_{\boldsymbol{\lambda}} \quad & F(\boldsymbol{\lambda}) = \sum_{i=1}^{m} \lambda_i, \\ s.t. \quad & L(\boldsymbol{\lambda}) = \frac{1}{2}\sum_{i=1}^{m} \lambda_i - 1 - \log \lambda_i \leq \epsilon, \qquad (d) \\ & U(\boldsymbol{\lambda}) = \frac{1}{2}\sum_{i=1}^{m} \frac{1}{\lambda_i} - 1 - \log \frac{1}{\lambda_i} \leq \delta, \quad (e) \\ & \lambda_i \geq 1, \ i = 1, \ldots, m. \end{aligned}$$

The convex property of Problem $\mathbf{P_4}$ is examined in the following proposition. The proof is given in the Appendix.

***Proposition** 1:* Problem $\mathbf{P_4}$ is a convex optimization problem when $\delta \leq \delta_0$, where $\delta_0 = -\frac{1}{4} - \frac{1}{2}\log\frac{1}{2}$. Otherwise, it is a non-convex problem.

***Proposition** 2:* The optimal solution of $\mathbf{P_4}$ makes at least one of the constraints (d) and (e) hold with equality.

Proof: Proposition 2 is true since in $\mathbf{P_4}$, both $L(\boldsymbol{\lambda})$ in (d) and $U(\boldsymbol{\lambda})$ in (e) are increasing with $\lambda_i \geq 1$ and the increase of each λ_i will enlarge the objective function $F(\boldsymbol{\lambda})$ of $\mathbf{P_4}$. ■

From Propositions 1 and 2, we have the following lemmas with the proof of Lemma 4 given in the Appendix.

***Lemma** 3 ([12], Theorem 7):* The optimal solution for Problem $\mathbf{P_4}$ with the constraint (e) removed is given by $F(\boldsymbol{\lambda})_{\max} = m\bar{\delta}(\frac{\epsilon}{m})$, when $\lambda_i = \bar{\delta}(\frac{\epsilon}{m})$ for $i = 1, \ldots, m$.

***Lemma** 4:* If $\delta \leq \delta_0$, then the optimal solution of Problem $\mathbf{P_4}$ without the constraint (d) is given by $F(\boldsymbol{\lambda})_{\max} = m/\underline{\delta}(\frac{\delta}{m})$, when $\lambda_i = 1/\underline{\delta}(\frac{\delta}{m})$ for $i = 1, \ldots, m$.

Based on Lemmas 2-4, we find that the largest performance degradation under (ϵ, δ)-stealthy attacks depends on the relationship between ϵ and δ, where either (d) or (e) in $\mathbf{P_4}$ can be neglected sometimes. By solving $\mathbf{P_4}$, the performance degradation under (ϵ, δ)-stealthy attacks is provided in Theorem 2.

***Theorem** 2:* Consider the system and detector above, and the information set Γ_1^∞ satisfying assumptions (A1)-(A3). For any (ϵ, δ)-stealthy attack $\tilde{u}_1^\infty$ that is generated by Γ_1^∞ and produces i.i.d. Gaussian innovation $\tilde{z}_k \sim \mathcal{N}(\mu, \Sigma_{\tilde{z}})$, the resulting estimation error of sensor measurements J satisfies

(1) If $\delta > D^\delta_{\max}(\epsilon)$, then

$$J \leq m\bar{\delta}(\frac{\epsilon}{m}) = \mathrm{tr}(PW) + m\left(\bar{\delta}(\frac{\epsilon}{m}) - 1\right) + \mathrm{tr}(\Sigma_z^{-1}R);$$

(2) If $\delta \leq \min\left(D^\delta_{\max}(\epsilon), \delta_0\right)$, then $J \leq m\underline{\delta}(\frac{\delta}{m})^{-1}$;

(3) If $\delta_0 < \delta \leq D^\delta_{\max}(\epsilon)$, then J is less than $F(\boldsymbol{\lambda})_{\max}$ in $\mathbf{P_4}$, where $\mathbf{P_4}$ is non-convex, but can be efficiently solved by monotonic optimization methods in [24].

Proof: Given an (ϵ, δ)-stealthy attack, the constraint (e) of Problem $\mathbf{P_4}$ can be removed if $\delta > D^\delta_{\max}(\epsilon)$ since it is satisfied definitely. As a consequence, by leveraging Lemma 3, the result of the first statement follows.

For the second statement, we will find the optimal solution of $\mathbf{P_4}$ by proving that the constraint (d) is always fulfilled and removable when $\delta \leq \min\left(D^\delta_{\max}(\epsilon), \delta_0\right) = \kappa$, and then using Lemma 4. Firstly, we divide $\delta \leq \kappa$ into two cases: $\delta \leq \min\left(D^\delta_{\min}(\epsilon), \kappa\right)$, and $\delta \in \left(D^\delta_{\min}(\epsilon), \kappa\right]$ if $D^\delta_{\min}(\epsilon) < \kappa$. For the former case where the inequality $\delta \leq D^\delta_{\min}(\epsilon)$ is true, Corollary 1 tells that $D^\epsilon_{\max}(\delta) \leq \epsilon$. Hence, the constraint (d) is always satisfied. Subsequently, for arbitrary δ in the latter case, there exists an ϵ_0 such that $\delta = D^\delta_{\max}(\epsilon_0)$. Based on Lemmas 3 and 4, the solution $\boldsymbol{\lambda}^*$ with $\lambda_i^* = \bar{\delta}(\frac{\epsilon_0}{m}) = \underline{\delta}(\frac{\delta}{m})^{-1}$ for $i = 1, \ldots, m$ is optimal for $\mathbf{P_4}$ which achieves both the equalities of (d) and (e) when the parameters are specified as (ϵ_0, δ). Further, since $\delta \leq D^\delta_{\max}(\epsilon)$ and the function $D^\delta_{\max}(x)$ is monotonically increasing on $x \geq 0$, we have $\epsilon \geq \epsilon_0$. Then, supposing that there exists a solution $\boldsymbol{\lambda}^0$ meeting $\epsilon_0 \leq L(\boldsymbol{\lambda}^0) \leq \epsilon$ and $U(\boldsymbol{\lambda}^0) = \delta$, we find $F(\boldsymbol{\lambda}^0) \leq F(\boldsymbol{\lambda}^*)$ since $\boldsymbol{\lambda}^*$ is already the optimal solution of $\mathbf{P_4}$ subject to the constraints $U(\boldsymbol{\lambda}) = \delta$ and $\lambda_i \geq 1, \ i = 1, \ldots, m$. Hence, $\mathbf{P_4}$ with (d) removed is equivalent to the original problem when $\delta \leq \kappa$. Consequently, by using Lemma 4, the result follows.

And for $\delta_0 < \delta \leq D^\delta_{\max}(\epsilon)$, Proposition 1 shows that $\mathbf{P_4}$ is non-convex. It is inherently difficult to provide the optimal solution of a non-convex problem in an analytical form. However, we find that in Problem $\mathbf{P_4}$, $F(\boldsymbol{\lambda}_1) \geq F(\boldsymbol{\lambda}_2)$ if $\boldsymbol{\lambda}_1 \geq \boldsymbol{\lambda}_2 > 0$ and so are $L(\boldsymbol{\lambda})$ and $U(\boldsymbol{\lambda})$, showing their monotonically increasing property. Based on these properties, $\mathbf{P_4}$ is a monotonic optimization problem according to the corresponding definition in Section 2.2 of [24], whose optimal solution can be approached by applying the Polyblock Outer Approximation algorithm in [24]. Further, since the objective function $F(\boldsymbol{\lambda})$ is Lipschitz continuous, the algorithm is guaranteed to converge to an η-optimal ($\eta > 0$) solution in a finite number of iterations. It means that the distance between the optimal solution $F(\boldsymbol{\lambda}^*)$ and the obtained one $F(\bar{\boldsymbol{\lambda}})$ is smaller than the given threshold η, i.e., $|F(\boldsymbol{\lambda}^*) - F(\bar{\boldsymbol{\lambda}})| < \eta$. ■

***Remark** 4:* When $\tilde{z}_k$ is i.i.d. Gaussian, Theorem 2 extends the results in [12], since for a given ϵ, our results coincide with that in [12] when $\delta \geq D^\delta_{\max}(\epsilon)$. For $\delta < D^\delta_{\max}(\epsilon)$, Theorem 2 gives a different upper bound of the performance degradation for (ϵ, δ)-stealthy attacks, which are actually ϵ-stealthy in [12].

C. Optimal Attacks

In this subsection, we design the optimal attacks that result in i.i.d. Gaussian innovations and achieve the upper bound of the performance degradation given in Theorem 2 when the system $\{A, B, C\}$ is right-invertible. The following tool lemma is given first, followed by the main results in Theorem 3.

***Lemma** 5 ([12], Lemma 10):* The system $\{A - KC, B, C\}$ is also right-invertible if the system $\{A, B, C\}$ is right-invertible.

***Theorem** 3:* Consider the problem setup for the right-invertible system $\{A, B, C\}$ in Section II and denote $\boldsymbol{\lambda}^*(\epsilon, \delta) = [\lambda_1^*, \ldots, \lambda_m^*]^T$ as the optimal or suboptimal solution of $\mathbf{P_4}$. Then, the (ϵ, δ)-stealthy attack $\tilde{u}_k = u_k + \pi_k$ achieves the largest performance degradation given in Theorem 2, where $\Lambda^* = \mathrm{diag}(\lambda_1^*, \ldots, \lambda_m^*)$ and π_k is the output

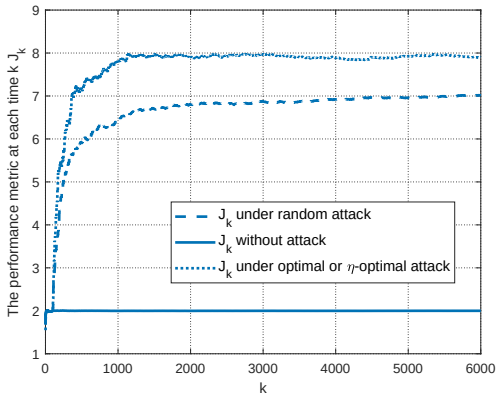

Fig. 2. The performance metric at each time slot J_k under the optimal or η-optimal attack and the randomly generated attack when $\epsilon = 2$ and $\delta = 0.54$.

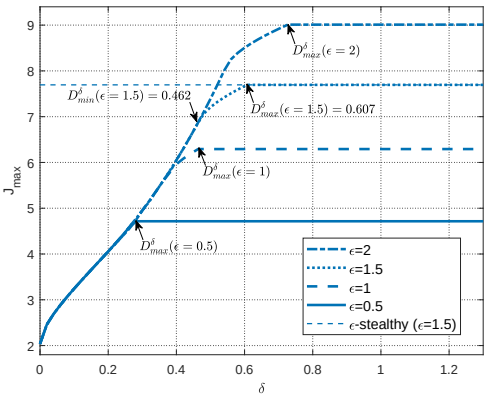

Fig. 3. The performance degradation induced by (ϵ, δ)-stealthy attacks with ϵ fixed and δ varying.

of the right inverse of the system (9) with i.i.d. sequence $\left\{\xi_1^\infty | \xi_k \sim \mathcal{N}\left(0, (\Lambda^* - I_m)\Sigma_z\right)\right\}$ as the input:

$$e_{k+1} = (A - KC)e_k + B\pi_k, \quad \xi_k = Ce_k, \tag{9}$$

where $\hat{x}_{k+1}^v = A\hat{x}_k^v + Kz_k^v + B\tilde{u}_k$, $e_k = \hat{x}_k^v - \hat{\tilde{x}}_k$ and $\xi_k = \tilde{z}_k - z_k^v$ with i.i.d. $z_k^v = \tilde{y}_k - C\hat{x}_k^v \sim \mathcal{N}(0, \Sigma_z)$.

Proof: The proof that the attack $\tilde{u}_k = u_k + \pi_k$ in Theorem 3 achieves the upper bound of the performance degradation given in Theorem 2 is similar to the results in [12, 14]. Hence, we omit this part of the proof due to the space limitation.

Next, we will show that the attack $\tilde{u}_k$ as described above is (ϵ, δ)-stealthy regardless of the relationship between ϵ and δ. Considering $\delta > D_{\max}^{\delta}(\epsilon)$, the attack $\tilde{u}_k$ is generated when $\lambda_i^* = \bar{\delta}(\frac{\epsilon}{m})$ for $i = 1, \ldots, m$. Using (7) and (8) and expanding $D_{KL}(\tilde{z}_k \| z_k)$ and $D_{KL}(z_k \| \tilde{z}_k)$ with λ_i^*'s, we have

$$D_{KL}(\tilde{z}_k \| z_k) = \frac{1}{2}\sum_{i=1}^{m} \lambda_i^* - 1 - \log \lambda_i^* = m \cdot \frac{\epsilon}{m} = \epsilon,$$

$$D_{KL}(z_k \| \tilde{z}_k) = \frac{1}{2}\sum_{i=1}^{m} \frac{1}{\lambda_i^*} - 1 + \log \lambda_i^* = D_{\max}^{\delta}(\epsilon) < \delta,$$

showing that the attack $\tilde{u}_k$ is (ϵ, δ)-stealthy.

For $\delta \leq \min\left(D_{\max}^{\delta}(\epsilon), \delta_0\right) = \kappa$, an optimal attack is given when $\lambda_i^* = 1/\underline{\delta}(\frac{\delta}{m})$ for $i = 1, \ldots, m$. Based on Corollary 1 and Theorem 2, we have $D_{KL}(z_k \| \tilde{z}_k) = m \cdot \frac{\delta}{m} = \delta$ and

$$D_{KL}(\tilde{z}_k \| z_k) = D_{\max}^{\epsilon}(\delta) \leq \epsilon, \text{ if } \delta \leq \min\left(D_{\min}^{\delta}(\epsilon), \kappa\right),$$

$$D_{KL}(\tilde{z}_k \| z_k) = \epsilon_0 \leq \epsilon, \text{ if } \delta \in \left(D_{\min}^{\delta}(\epsilon), \kappa\right], \bar{\delta}(\frac{\epsilon_0}{m}) = 1/\underline{\delta}(\frac{\delta}{m}).$$

As for $\delta_0 < \delta \leq D_{\max}^{\delta}(\epsilon)$, $\boldsymbol{\lambda}^*(\epsilon, \delta)$ is obtained by solving Problem $\mathbf{P_4}$, for which the generated attack $\tilde{u}_k$ must be (ϵ, δ)-stealthy. At this point, the proof is complete. ■

V. Numerical Results

In this section, we demonstrate our results by conducting several numerical simulations on the system with parameters

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, B = I_2, C = \begin{bmatrix} 3 & 4 \\ 1 & 1 \end{bmatrix}, Q = R = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.3 \end{bmatrix},$$

which is a right-invertible system and used in [14]. We firstly illustrate the superiority of the optimal or η-optimal attack in Fig. 2. The parameters are selected as $\epsilon = 2$ and $\delta = 0.54$ and the results in the figure are obtained by conducting 1000 times simulations. During the time interval $[0, 100]$, the Kalman filter equipped in the system evolves into a steady-state. Subsequently, the optimal or suboptimal attack (dotted line) and the randomly generated attack (dashed line) are applied, respectively, wherein the latter one we make at least one of the constraints (d) and (e) in $\mathbf{P_4}$ hold with equality. Obviously, the attack obtained based on Theorems 2 and 3 outperforms others. In addition, the convergence value of J_k under the optimal or suboptimal attack (about 8) coincides with the result in Theorem 2 (see Fig. 3).

Then, we analyze the performance degradation on the estimation error of sensor measurements induced by (ϵ, δ)-stealthy attacks with various stealth levels. The results of the upper bound of the performance degradation $J_{\max}$ given in Theorem 2 are presented in Fig. 3, where the ϵ in each curve is fixed to a different value. As observed from the figure, $J_{\max}$ is increasing with δ for a given ϵ until it reaches $m\bar{\delta}(\frac{\epsilon}{m})$, which occurs because the constraint brought by the parameter δ does not contribute to restricting the attack space when δ exceeds $D_{\max}^{\delta}$. Additionally, applying attacks with a larger ϵ does not necessarily means a higher $J_{\max}$, at least not lower, but it would result in a larger performance loss eventually as δ increases. The attacks with different ϵ have the same $J_{\max}$ when δ is small, since it is the parameter δ that serves to limit the attack space. With δ increasing, ϵ begins to constrain the feasible attack space. As a result, the attack with a larger ϵ potentially has a larger $J_{\max}$ since a broader attack space can be provided.

To compare our proposed attack with the ϵ-stealthy attack in [11, 12] when the corrupted innovation is an i.i.d. Gaussian process, we take $\epsilon = 1.5$ (dotted line and thin dashed line in Fig. 3) as an example. Since ϵ-stealthiness does not rely on δ, the ϵ-stealthy attack has a constant $J_{\max}$, which coincides with that of (ϵ, δ)-stealthy attack when $\delta > D_{\max}^{\delta}(\epsilon)$. When the specified parameter $\delta \in \left[D_{\min}^{\delta}(\epsilon), D_{\max}^{\delta}(\epsilon)\right]$, the (ϵ, δ)-stealthy attacks have lower $J_{\max}$ though they are of the same ϵ-stealthiness, indicating that the (ϵ, δ)-stealthiness can further refine the ϵ-stealthiness. And for $\delta < D_{\min}^{\delta}(\epsilon)$, the maximum achievable exponent of p_k^F is smaller than ϵ. Strictly speaking,

the (ϵ, δ)-stealthy attack is not ϵ-stealthy now, but ϵ_0-stealthy where $\epsilon_0 < \epsilon$ and $\bar{\delta}(\frac{\epsilon_0}{m}) = \underline{\delta}(\frac{\delta}{m})^{-1}$, and thus its performance degradation is smaller than that of all ϵ-stealthy attacks. In summary, considering the system performance degradation, our proposed notion describes the extent of attack stealthiness in a more fine-grained way, especially when the given parameters satisfy $D^{\delta}_{\min}(\epsilon) \leq \delta \leq D^{\delta}_{\max}(\epsilon)$.

VI. Conclusion

In this technical note, we have investigated the stealthiness and performance of stealthy actuator signal attacks in stochastic cyber-physical systems with detection algorithms unknown. To quantify the detectability of attacks, a notion of (ϵ, δ)-stealthiness has been proposed, which considers the convergence rates of false alarm probability and detection probability. Then, we have studied the attacks that produce i.i.d. Gaussian innovations and characterized the largest performance degradation on the estimation of sensor measurements induced by (ϵ, δ)-stealthy attacks based on the tradeoff between the largest exponent of the two probabilities in detectors. In addition, the attack achieving the largest performance loss in right-invertible systems has been designed. Based on our analysis and simulations, the (ϵ, δ)-stealthiness better characterizes the extent of attack stealthiness.

Appendix

Proof of Proposition 1: It is equivalent for Problem $\mathbf{P_4}$ to change its objective to $\min_{\boldsymbol{\lambda}} -F(\boldsymbol{\lambda}) = \sum_{i=1}^{m} -\lambda_i$, which is convex. For the constraint (d), it is explicitly convex due to the convexity of $\lambda_i - 1 - \log \lambda_i$ for all $\lambda_i \geq 1$. Since the function $\frac{1}{\lambda_i} + \log \lambda_i - 1$ is convex on the interval $\lambda_i \in [1, \frac{1}{\underline{\delta}(\delta)}]$ if $\delta \leq \delta_0$ where $\delta_0 = -\frac{1}{4} - \frac{1}{2}\log\frac{1}{2}$, the constraint (e) is also convex. Otherwise, the constraint (e) is non-convex. Consequently, $\mathbf{P_4}$ is a convex optimization problem only when $\delta \leq \delta_0$.

And we resort to a contradiction for the proof of the infeasibility of translating $\mathbf{P_4}$ into a convex form. In another word, we will show that the contraint (e) is never convex if $\delta > \delta_0$ while keeping the objective function convex. Assume that there exists a function $g(\boldsymbol{\varphi}) = \sum_{i=1}^{m} g_i(\varphi_i) = -F(\boldsymbol{\lambda})$ with $g_i(\varphi_i) = -\lambda_i \in [-\frac{1}{\underline{\delta}(\delta)}, -1]$ where $g_i(\cdot)$ is convex on φ_i. Note that $\boldsymbol{\varphi} = [\varphi_1, \ldots, \varphi_m]^T$. Hence, the convexity of the objective function $g(\boldsymbol{\varphi})$ still holds. Define $h_i(\varphi_i) = -1/g_i(\varphi_i) + \log(-g_i(\varphi_i)) - 1$, and its second derivative is

$$\begin{aligned} h_i''(\varphi_i) = &- g_i'(\varphi_i)^2 g_i(\varphi_i)^{-2}\left(2g_i(\varphi_i)^{-1} + 1\right) \\ &+ g_i''(\varphi_i) g_i(\varphi_i)^{-1}\left(g_i(\varphi_i)^{-1} + 1\right). \end{aligned}$$

Since the inequalities $g_i''(\varphi_i) \geq 0$, $g_i(\varphi_i)^{-1} \leq 0$ and $g_i(\varphi_i)^{-1} + 1 \geq 0$ hold due to the convexity and range of $g_i(\varphi_i)$, we have $h_i''(\varphi_i) \leq 0$ on the interval $-\frac{1}{\underline{\delta}(\delta)} \leq g_i(\varphi_i) \leq -\frac{1}{\underline{\delta}(\delta_0)}$ if $\delta > \delta_0$. Then, denoting the set of constraint (e) as $\Phi' = \{\boldsymbol{\varphi} | \sum_{i=1}^{m} h_i(\varphi_i) \leq \delta\}$, we can easily find that Φ' is non-convex. Therefore, the proof is complete. ∎

Proof of Lemma 4: From Propositions 1 and 2, if $\delta \leq \delta_0$, Problem $\mathbf{P_4}$ without the constraint (d) can be viewed as

$$\max \sum_{i=1}^{m} \underline{\delta}(\tau_i)^{-1}, \ s.t. \ \sum_{i=1}^{m} \tau_i = \delta, \tau_i \geq 0, i = 1, \ldots, m,$$

if we let $\lambda_i = \underline{\delta}(\tau_i)^{-1}$, where $\underline{\delta}(\tau_i)^{-1}$ is concave on $0 \leq \tau_i \leq \delta \leq \delta_0$. Subsequently, based on Jensen's inequality, the optimal solution of $\mathbf{P_4}$ is obtained as $F(\boldsymbol{\lambda})_{\max} = m\underline{\delta}(\frac{\delta}{m})^{-1}$ when $\tau_i = \frac{\delta}{m}$ for $i = 1, \ldots, m$. ∎

References

[1] Y. Mo, R. Chabukswar, and B. Sinopoli. Detecting integrity attacks on SCADA systems. *IEEE Transactions on Control Systems Technology*, 22(4):1396–1407, 2014.

[2] F. Pasqualetti, F. Dörfler, and F. Bullo. Attack detection and identification in cyber-physical systems. *IEEE Transactions on Automatic Control*, 58(11):2715–2729, 2013.

[3] S. Karnouskos. Stuxnet worm impact on industrial cyber-physical system security. In *Proceedings of the 37th IEEE Annual Conference on Industrial Electronics Society (IECON'2011)*, pages 4490–4494, 2011.

[4] J. Slay and M. Miller. Lessons learned from the Maroochy water breach. In *Proceedings of the 2007 International Conference on Critical Infrastructure Protection*, pages 73–82. Springer, 2007.

[5] Y. Chen, S. Kar, and J. Moura. Optimal attack strategies subject to detection constraints against cyber-physical systems. *IEEE Transactions on Control of Network Systems*, 5(3):1157–1168, 2018.

[6] Y. Chen, S. Kar, and J. Moura. Cyber-physical attacks with control objectives. *IEEE Transactions on Automatic Control*, 63(5):1418–1425, 2018.

[7] Y. Mo and B. Sinopoli. Integrity attacks on cyber-physical systems. In *Proceedings of the 1st International Conference on High Confidence Networked Systems*, pages 47–54, 2012.

[8] Y. Liu, P. Ning, and M. K. Reiter. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security*, 14(1):13, 2011.

[9] D. I. Urbina, J. A. Giraldo, A. A. Cardenas, N. O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, and H. Sandberg. Limiting the impact of stealthy attacks on industrial control systems. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS'2016)*, pages 1092–1105, 2016.

[10] S. Sundaram and C. N. Hadjicostis. Distributed function calculation via linear iterative strategies in the presence of malicious agents. *IEEE Transactions on Automatic Control*, 56(7):1495–1508, 2011.

[11] C.-Z. Bai, F. Pasqualetti, and V. Gupta. Security in stochastic control systems: Fundamental limitations and performance bounds. In *Proceedings of the 2015 American Control Conference*, pages 195–200, 2015.

[12] C.-Z. Bai, F. Pasqualetti, and V. Gupta. Data-injection attacks in stochastic control systems: Detectability and performance tradeoffs. *Automatica*, 82:251–260, 2017.

[13] C.-Z. Bai, V. Gupta, and F. Pasqualetti. On Kalman filtering with compromised sensors: Attack stealthiness and performance bounds. *IEEE Transactions on Automatic Control*, 62(12):6641–6648, 2017.

[14] E. Kung, S. Dey, and L. Shi. The performance and limitations of ϵ-stealthy attacks on higher order systems. *IEEE Transactions on Automatic Control*, 62(2):941–947, 2017.

[15] R. Zhang and P. Venkitasubramaniam. Stealthy control signal attacks in linear quadratic gaussian control systems: Detectability reward tradeoff. *IEEE Transactions on Information Forensics and Security*, 12(7):1555–1570, 2017.

[16] Z. Guo, D. Shi, K. H. Johansson, and L. Shi. Worst-case stealthy innovation-based linear attack on remote state estimation. *Automatica*, 89:117–124, 2018.

[17] Y. Polyanskiy and Y. Wu. *Lecture Notes on Information Theory*. MIT (6.441), UIUC (ECE 563), 2012-2016.

[18] L. Garcia, F. Brasser, M. H. Cintuglu, A. R. Sadeghi, O. A. Mohammed, and S. A. Zonouz. Hey, my malware knows physics! attacking plcs with physical model aware rootkit. In *Proceedings of the 2017 Network and Distributed System Security Symposium (NDSS'2017)*, 2017.

[19] S. Kullback. *Information Theory and Statistics*. Courier Dover, 1997.

[20] M. Wei, X. Cheng, and Q. Wang. A canonical decomposition of the right invertible system with applications. *SIAM Journal on Matrix Analysis and Applications*, 31(4):1958–1981, 2010.

[21] B. D. Anderson and J. B. Moore. *Optimal Filtering*. Courier Corporation, 2012.

[22] H. V. Poor. *An Introduction to Signal Detection and Estimation*. Springer Science & Business Media, 2013.

[23] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 2015.

[24] Y. J. A. Zhang, L. Qian, and J. Huang. Monotonic optimization in communication and networking systems. *Foundations and Trends® in Networking*, 7(1):1–75, 2013.