Automatica 136 (2022) 110079

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

How vulnerable is innovation-based remote state estimation: Fundamental limits under linear attacks☆

Hanxiao Liu [a,b], Yuqing Ni [c], Lihua Xie [a,*], Karl Henrik Johansson [b]

[a] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[b] School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden
[c] School of Internet of Things Engineering, Jiangnan University, Wuxi, China

ARTICLE INFO

Article history:
Received 7 September 2020
Received in revised form 4 October 2021
Accepted 14 October 2021
Available online 10 December 2021

ABSTRACT

This paper is concerned with the problem of how secure the innovation-based remote state estimation can be under linear attacks. A linear time-invariant system equipped with a smart sensor is studied. A metric based on Kullback–Leibler divergence is adopted to characterize the stealthiness of the attack. The adversary aims to maximize the state estimation error covariance while stay stealthy. The maximal performance degradations that an adversary can achieve with any linear first-order false-data injection attack under strict stealthiness for vector systems and ϵ-stealthiness for scalar systems are characterized. We also provide an explicit attack strategy that achieves this bound and compare this attack strategy with strategies previously proposed in the literature. Finally, some numerical examples are given to illustrate the results.

1. Introduction

Cyber–physical systems (CPS), which closely integrate computational elements and physical processes, are playing a critical role in society. Any successful cyber–physical attack can bring huge damages to critical infrastructure or even human lives. Maroochy water breach in 2000 (Slay & Miller, 2007), Stuxnet malware in 2010 (Karnouskos, 2011), Ukraine power outage in 2015 (Whitehead et al., 2017), Venezuela blackouts in 2019 (Jones, 2019) are examples of incidents that motivate us to pay more attention to the security of CPS.

An adversary may launch attacks to disturb the monitoring and state estimation of CPS. Many existing works focus on designing detection algorithms and secure state estimation strategies to enhance the security of CPS. Mo et al. (Mo et al., 2013, Mo & Sinopoli, 2009, Mo et al., 2015) analyzed the effect of replay attacks where the attackers do not know the system information but replay the recorded measurements. They proposed a physical watermarking scheme to detect the attack. An algorithm that employs a satisfiability modulo theory paradigm was proposed in Shoukry et al. (2017) to tackle the complexity of secure state estimation. Teixeira et al. (2012) characterized the properties of zero dynamics attacks and provided necessary and sufficient conditions for which input and output deviations should satisfy to reveal attacks. A secure state estimation algorithm was presented in Mishra et al. (2016), and upper bounds on the state estimation error covariance, when the maximum number of attacked sensors is known, were derived.

The problem of what is the worst possible attack is of great interest to help in the search for defence strategies. Mo and Sinopoli (2015) formulated a constrained control problem subject to the attacker's strategy and characterized its maximal perturbation. A linear quadratic function was employed to capture the attacker's control goal and constraints in Chen et al. (2017). The authors stated that linear feedback is the optimal attack strategy and provided two algorithms to derive the optimal attack sequence. In Zhang et al. (2015), the problem of scheduling a denial-of-service (DoS) attack with limited energy was studied. The optimal attack schedule in a special scenario was proposed and the optimal attack schedule with both energy constrained sensor and attacker was analyzed. A similar problem but with a packet-dropping network was studied in Qin et al. (2018).

To the best of our knowledge, the concept of stealthiness of the attack was first introduced as δ-marginal stealthiness (δ-MS) in Bai and Gupta (2014). The authors characterized a stealthiness level from the probability of false alarm and investigated

☆ This work is supported by the A*STAR Industrial Internet of Things Research Program under the RIE2020 IAF-PP Grant A1788a0023, Singapore, the Knut and Alice Wallenberg Foundation, Sweden, the Swedish Foundation for Strategic Research, and the Swedish Research Council. The material in this paper was partially presented at the 21st IFAC World Congress (IFAC 2020), July 12–17, 2020, Berlin, Germany. This paper was recommended for publication in revised form by Associate Editor Luca Schenato under the direction of Editor Christos G. Cassandras.

* Corresponding author.
E-mail addresses: hanxiao001@ntu.edu.sg (H. Liu), yuqingni@jiangnan.edu.cn (Y. Ni), elhxie@ntu.edu.sg (L. Xie), kallej@kth.se (K.H. Johansson).

https://doi.org/10.1016/j.automatica.2021.110079
0005-1098/

the trade-off between the performance degradation of the state estimation and the stealthiness level. Based on this work, a notion of ϵ-stealthiness based on Kullback–Leibler (KL) divergence to quantify attack detectability was proposed and the maximal performance degradation under ϵ-stealthy attack strategy was revealed in Bai et al. (2015), Bai, Pasqualetti, et al. (2017). The authors of Kung et al. (2016) generalized the above results to vector systems. Furthermore, Bai, Gupta, et al. (2017) was devoted to seeking the optimal attack by compromising sensor measurements. In this paper, we adopt the same stealthiness metric as in Bai, Gupta, et al. (2017), Bai et al. (2015). Different from these works focusing on designing the optimal attack strategy after deriving the performance degradation bound, in this paper we obtain the maximal performance loss under linear attacks.

Innovation-based linear integrity attacks were first studied in Guo et al. (2016). An optimal linear attack policy was proposed to achieve the maximal performance degradation while not being detected by a χ^2 detector, which can also be considered as a strictly stealthy attack as proposed in Bai and Gupta (2014). Some extensions of this work can be found in Guo et al. (2017), Guo et al. (2019). These authors also investigated this type of attacks in the detection framework based on KL divergence (Guo et al., 2018). Different from previous attack strategies consisting of a zero-mean random variable, a more general linear attack strategy with an arbitrary mean random variable was studied in Li and Yang (2019). However, all the above papers only consider memory less attacks. A larger performance degradation of the remote estimator can be expected when the attacker utilizes both past and present information. Motivated by this point, we consider how vulnerable innovation-based remote state estimators are to a linear attack which leverages both past and present innovation. Moreover, we allow for sequence detection instead of just one-slot detection.

The main contributions of this paper are as follows:

1. A fundamental performance degradation bound is provided for innovation-based remote state estimators under strictly stealthy linear attacks (Theorem 3.1). A worst-case attack strategy achieving the bound is explicitly stated.
2. For ϵ-stealthy linear attacks, the corresponding degradation bound is derived for scalar systems (Theorem 4.1). Again, a worst-case attack strategy achieving the bound is explicitly stated.
3. The proposed attack strategies are shown to outperform other attack strategies discussed in the literature. It is illustrated how the memory in the proposed strategy provides specific advantages from an adversary point of view.

Some preliminary results are described in our conference paper (Liu et al., 2020). The main differences between the current paper and Liu et al. (2020) are significant: (1) The attack model proposed in this paper is more general. (2) We generalize the results of strictly stealthy attacks to vector systems and provide worst performance degradation ratio for the estimation error covariance. (3) Detailed proofs of theorems and lemmas are included. (4) Simulations to validate our theoretical findings are provided.

The rest of the paper is organized as follows. Section 2 formulates the problem by introducing the system model, attack model as well as two stealthiness metrics. We present the worst-case performance degradation bounds for remote state estimation under strictly stealthy attacks for vector systems and ϵ-stealthy attacks for scalar systems in Sections 3 and 4, respectively. In Section 5, some numerical examples are provided to verify the performance of the proposed strategies and compare them with strategies from the literature. Conclusions are provided in Section 6. For the sake of readability, some proofs are included in the appendix.

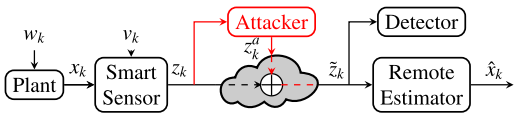


Fig. 1. The system diagram.

Notations:

The notation $x_{k_1}^{k_2}$ is the sequence $\{x_{k_1}, x_{k_1+1}, \ldots, x_{k_2}\}$. The spectral radius is defined as $\rho(A) \triangleq \max\{|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n|\}$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the matrix $A \in \mathbb{R}^{n\times n}$. I_n denotes the identity matrix of order n. The zero matrix $\mathbf{0}_{m\times n}$ is the $m \times n$ matrix with all entries equal to 0. The transpose of matrix A is represented by $A^\intercal$. $\mathbb{S}_+^n$ ($\mathbb{S}_{++}^n$) is the set of $n \times n$ positive semi-definite (definite) matrices. When $X \in \mathbb{S}_+^n$ ($\mathbb{S}_{++}^n$), we simply write $X \succeq 0$ ($X \succ 0$).

2. Problem formulation

In this section, the system and attack models are introduced together with the stealthiness and performance degradation metrics. Finally, the problem of interest is formulated. The diagram for the considered system is illustrated in Fig. 1. A smart sensor measures the output of a physical plant and transmits the innovation to a remote estimator via a wireless network. An attacker attempts to modify the transmission data, which are received by a remote estimator and a detector. The detailed system model is presented in this section.

2.1. System model

Consider a linear time-invariant (LTI) system described by the following equations:

$$x_{k+1} = Ax_k + w_k, \tag{1}$$

$$y_k = Cx_k + v_k, \tag{2}$$

where $x_k \in \mathbb{R}^n$ and $y_k \in \mathbb{R}^p$ are the vector of state variables and sensor measurements at time k, respectively. $w_k \in \mathbb{R}^n$ denotes the process noise and $v_k \in \mathbb{R}^p$ the measurement noise. They are assumed to be mutually independent zero-mean Gaussian variables with covariances $Q \succeq 0$ and $R \succ 0$, i.e., $w_k \sim \mathcal{N}(0, Q)$ and $v_k \sim \mathcal{N}(0, R)$. We further assume that x_0 is a zero mean Gaussian random vector independent of the process noise and the measurement noise, and with covariance Σ. We focus on stable systems.

Assumption 2.1. The spectral radius $\rho(A) < 1$.

The system is equipped with a local smart sensor whose functions include signal conditioning, signal processing, and decision making (Lewis, 2004). In our work, the smart sensor employs the Kalman filter to process measurement and transmit the innovation to the remote estimator:

$$\hat{x}_{k+1|k}^s = A\hat{x}_k^s,$$
$$P_{k+1|k} = AP_{k|k}A^\intercal + Q,$$
$$K_k = P_{k|k-1}C^\intercal(CP_{k|k-1}C^\intercal + R)^{-1},$$
$$\hat{x}_k^s = \hat{x}_{k|k-1}^s + K_k(y_k - C\hat{x}_{k|k-1}^s),$$
$$P_{k|k} = P_{k|k-1} - K_kCP_{k|k-1},$$

with initialization $\hat{x}_{0|-1} = x_0$.

Under Assumption 2.1, the Kalman gain will converge exponentially. Therefore, we consider a steady-state Kalman filter with gain K and a priori minimum mean square error (MMSE) P:

$$P \triangleq \lim_{k\to\infty} P_{k|k-1}, \tag{3}$$

$$K \triangleq PC^{\intercal}(CPC^{\intercal}+R)^{-1}. \tag{4}$$

As a result, the Kalman filter can be rewritten as:

$$\hat{x}^s_{k+1|k} = A\hat{x}^s_k,\ \hat{x}^s_k = \hat{x}^s_{k|k-1} + Kz_k,$$

where $z_k \triangleq y_k - C\hat{x}^s_{k|k-1}$ is the innovation at time k, which is transmitted to the remote estimator. Recall that $z_k \sim \mathcal{N}(0, \Sigma_z)$, where $\Sigma_z \triangleq CPC^{\intercal}+R \succ 0$. Since $y_k = z_k + C\hat{x}^s_{k|k-1}$, one can argue that z_k contains the same information about the uncertainty in the process as y_k. It is worth noticing that transmitting the raw sensor measurement does not make the system safer (Guo et al., 2019). In the literature (Guo et al., 2016, Guo et al., 2019, Li et al., 2017, Ribeiro et al., 2006), similar setups have been considered.

2.2. Attack model

The adversary is assumed to have the following capabilities:

1. The attacker has access to all innovations from the smart sensor, i.e., it knows the innovations $z_1, \ldots, z_k$ at time k.
2. The attacker can modify the innovations to arbitrary values.
3. The attacker has knowledge of the system matrix A, the measurement matrix C, as well as the covariances, i.e., Q and R, of the noises.

Remark 2.1. The third capability can be relaxed. If the attacker does not have access to A but it can access the input and output, it can identify the system parameters. The accuracy of the identification will affect the attack performance. This will be illustrated in Section 5.

The attacker injects the false data z^a_k and modifies the innovations in real-time as:

$$\tilde{z}_k = T\tilde{z}_{k-1} + Sz_k + \phi_k, \tag{5}$$

where $T \in \mathbb{R}^{p\times p}$ and $S \in \mathbb{R}^{p\times p}$ are matrices to be chosen by the attacker, and $\phi_k \sim \mathcal{N}(0, \Phi)$ is an i.i.d. Gaussian random variable with covariance $\Phi \in \mathbb{S}^p_+$, which is independent of z_k. The attack model (5) suggests that the attacker can generate the false-data injection attack based on filtering the innovation sequence from the smart sensor with a linear type potentially driven by noise.

Remark 2.2. Observe that the works (Guo et al., 2016, 2017, 2018, Guo et al., 2019) consider memoryless attacks, i.e., the attack is only based on the current innovation. Here, we seek to explore the possibility of a larger performance degradation for the remote estimator when the attacker utilizes both past and present information. More specifically, we focus on a linear time-invariant first order attack model and characterize the maximal performance degradation that an adversary can achieve. We also provide an explicit attack strategy that achieves this bound. It is hoped that our work can provide some insight into other more general attack models such as linear-time varying and nonlinear attack models.

The remote estimator receives $\tilde{z}_k$ so the remote state estimation follows:

$$\hat{x}_{k|k-1} = A\hat{x}_{k-1}, \tag{6}$$

$$\hat{x}_k = \hat{x}_{k|k-1} + K\tilde{z}_k. \tag{7}$$

Here, we initialize $\hat{x}_{1|0} = \hat{x}^s_{1|0}$ and $\tilde{z}_k = 0$ when $k \le 0$.

2.3. Detector and stealthiness metric

The attacker wants to be stealthy, otherwise the system will take countermeasures to keep a safe operation. We employ a metric based on KL divergence to quantify stealthiness, as first proposed in Bai et al. (2015).

The attack detection problem is posed as sequential hypothesis testing. The detector uses the received sequence to carry out the following binary hypothesis testing:

H_0 : There is no attack in process. (The remote estimator receives z^k_1).

H_1 : There is an attack in process. (The remote estimator receives $\tilde{z}^1_k$).

In testing H_0 versus H_1 there are two types of possible errors: the first type is called "false alarm", which denotes that the detector decides H_1 given H_0, and the second type is called "miss detection", which represents that the detector decides H_0 when H_1 is correct. Here, we denote the probability of miss detection at time k as p^M_k, and the probability of false alarm as p^F_k. Furthermore, the probability of correct detection is p^D_k, which denotes that the detector decides H_1 given H_1. Obviously, $p^D_k + p^M_k = 1$. We provide two definitions for attack stealthiness:

Definition 2.1 (*Strictly Stealthy Attack (Bai, Gupta, et al., 2017)*)**.** The attack is strictly stealthy if $p^F_k \ge p^D_k$, $(k \ge 0)$, holds for any detector.

Definition 2.2 (ϵ*-stealthy Attack Bai, Gupta, et al., 2017*[1]**).** The attack is ϵ-stealthy if

$$\limsup_{k\to\infty} -\frac{1}{k}\log p^F_k \le \epsilon \tag{8}$$

holds for any detector that satisfies $0 < p^M_k < \delta$ for all $k \ge 0$, where $0 < \delta < 1$.

2.4. Performance degradation metric

We employ the ratio of the trace of the covariance of the state estimation error $\tilde{P}$ and P to quantify the performance degradation introduced by the attacker, i.e.,

$$\eta = \frac{\operatorname{tr}\tilde{P}}{\operatorname{tr}P}, \tag{9}$$

where P is defined in (3) and $\tilde{P}$ is defined as follows:

$$\tilde{P} \triangleq \limsup_{k\to\infty} \frac{1}{k}\sum_{l=1}^{k} \tilde{P}_l, \tag{10}$$

where $\tilde{P}_l = E[(x_l - \hat{x}_{l|l-1})(x_l - \hat{x}_{l|l-1})^{\intercal}]$. When there is no attack, $\tilde{z}_k = z_k$. As $\hat{x}_{1|0} = \hat{x}^s_{1|0}$, one can derive that $\hat{x}_{k|k-1} = \hat{x}^s_{k|k-1}$. Hence, $\tilde{P} = P$ and $\eta = 1$. In other words, the performance will not be degraded without attacks.

2.5. KL divergence

In order to quantify the stealthiness level of attacks, we need to employ the KL divergence (Cover & Thomas, 2012, Kullback & Leibler, 1951), which is defined as:

[1] Motivated by the Chernoff–Stain Lemma, the notion of ϵ-stealthiness was first proposed in Bai, Gupta, et al. (2017).

Definition 2.3 (*KL Divergence*). Let $\tilde{z}_1^k$ and z_1^k be two random sequences with joint probability density functions $f_{\tilde{z}_1^k}$ and $f_{z_1^k}$, respectively. The KL divergence between $\tilde{z}_1^k$ and z_1^k equals

$$D(\tilde{z}_1^k \parallel z_1^k) = \int_{-\infty}^{+\infty} \log \frac{f_{\tilde{z}_1^k}(\xi_1^k)}{f_{z_1^k}(\xi_1^k)} f_{\tilde{z}_1^k}(\xi_1^k) d\xi_1^k. \tag{11}$$

One can see that $D(\tilde{z}_1^k \parallel z_1^k) \geq 0$, and $D(\tilde{z}_1^k \parallel z_1^k) = 0$ if and only if $f_{\tilde{z}_1^k} = f_{z_1^k}$. Generally, KL divergence is asymmetric, i.e., $D(\tilde{z}_1^k \parallel z_1^k) \neq D(z_1^k \parallel \tilde{z}_1^k)$.

Necessary and sufficient conditions for strictly stealthy attacks and ϵ-stealthy attacks are provided in Bai, Gupta, et al. (2017):

Lemma 2.1 (*Strictly Stealthy Attacks Bai, Gupta, et al., 2017*). *An attack sequence $\tilde{z}_1^k$ is strictly stealthy if and only if $\{\tilde{z}_1, \tilde{z}_2, \ldots\}$ is a sequence of i.i.d. Gaussian random variables with zero mean and covariance* $\mathrm{Cov}(\tilde{z}_k) = \Sigma_z = CPC^\intercal + R$.

Lemma 2.2 (ϵ-*stealthy Attacks Bai, Gupta, et al., 2017*). *If an attack $\tilde{z}_1^k$ is ϵ-stealthy, then*

$$\limsup_{k\to\infty} \frac{1}{k} D(\tilde{z}_1^k \parallel z_1^k) \leq \epsilon. \tag{12}$$

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

$$\lim_{k\to\infty} \frac{1}{k} D(\tilde{z}_1^k \parallel z_1^k) \leq \epsilon, \tag{13}$$

then the attack is ϵ-stealthy.

2.6. Problem of interest

We aim to derive fundamental vulnerabilities of innovation-based remote estimation. In other words, we seek to obtain how secure one can make innovation-based remote state estimation under linear attacks.

Given the innovation-based remote state estimator system in Fig. 1, how vulnerable is such a system under attack (5)? The answer is given by the worst performance degradation (9). For strictly stealthy and ϵ-stealthy attacks, these degradation can be formulated as the following optimization problems:

1. The attack is strictly stealthy:

$$\begin{aligned} \arg\max_{T,S,\Phi} \quad & \eta_s \triangleq \limsup_{k\to\infty} \frac{\frac{1}{k}\sum_{l=1}^{k} \mathrm{tr}\, \tilde{P}_l}{\mathrm{tr}\, P}, \\ \text{s.t.} \quad & \text{attack is strictly stealthy.} \end{aligned} \tag{14}$$

2. The attack is ϵ-stealthy:

$$\begin{aligned} \arg\max_{T,S,\Phi} \quad & \eta_\epsilon \triangleq \limsup_{k\to\infty} \frac{\frac{1}{k}\sum_{l=1}^{k} \mathrm{tr}\, \tilde{P}_l}{\mathrm{tr}\, P}, \\ \text{s.t.} \quad & \text{attack is } \epsilon\text{-stealthy.} \end{aligned} \tag{15}$$

We seek to obtain the optimal attack tuple (T^*, S^*, Φ^*) to maximize the performance degradation, while guaranteeing the prespecified stealthiness level.

3. Strictly stealthy attacks

The following theorem characterizes the maximal performance degradation ratio under a strictly stealthy attack. We also specify the optimal attack strategy.

For the simplicity of notations, we define

$$\mathcal{P}_1 \triangleq K \Sigma_z K^\intercal.$$

Theorem 3.1. *Consider system (1)–(2) satisfying Assumption 2.1. For strictly stealthy attacks of the form (5), the worst performance degradation ratio for the estimation error covariance is*

$$\eta_s = 1 + 4\frac{\mathrm{tr}\, \mathcal{X}}{\mathrm{tr}\, P},$$

where $\mathcal{X} = \mathcal{X}_1 - \mathcal{P}_1$ and $\mathcal{X}_1$ is the solution to the Lyapunov equation: $\mathcal{X}_1 = A\mathcal{X}_1 A^\intercal + \mathcal{P}_1$. The corresponding attack strategy is $(T^*, S^*, \Phi^*) = (\mathbf{0}_{m\times m}, -I_m, \mathbf{0}_{m\times m})$.

Proof. Rewrite (5) as follows:

$$\begin{aligned} \tilde{z}_l &= T\tilde{z}_{l-1} + Sz_l + \phi_l \\ &= \sum_{i=1}^{l} T^{l-i} S z_i + \sum_{i=1}^{l} T^{l-i}\phi_i. \end{aligned} \tag{16}$$

By Lemma 2.1, the covariance of $\tilde{z}_l$ ($l = 1, 2, \ldots$) needs to satisfy

$$\mathrm{Cov}(\tilde{z}_l) = \sum_{i=0}^{l-1} T^i \left(S\Sigma_z S^\intercal + \Phi\right)(T^i)^\intercal = \Sigma_z.$$

The feasible solution thus is $(T, S, \Phi) = (\mathbf{0}_{m\times m}, S, \Sigma_z - S\Sigma_z S^\intercal)$, where $\Sigma_z - S\Sigma_z S^\intercal \geq 0$. i.e., $\tilde{z}_l = Sz_l + \phi_l$, and the covariance of ϕ_l is $\Sigma_z - S\Sigma_z S^\intercal$. Now we derive the ratio η_s. Let us rewrite $\tilde{P}_l$ as follows:

$$\begin{aligned} \tilde{P}_l =& \mathbb{E}[(x_l - \hat{x}_{l|l-1})(x_l - \hat{x}_{l|l-1})^\intercal] \\ =& \mathbb{E}[(x_l - \hat{x}^s_{l|l-1})(x_l - \hat{x}^s_{l|l-1})^\intercal] \\ &+ \mathbb{E}[(\hat{x}^s_{l|l-1} - \hat{x}_{l|l-1})(\hat{x}^s_{l|l-1} - \hat{x}_{l|l-1})^\intercal] \\ &+ 2\mathbb{E}[(x_l - \hat{x}^s_{l|l-1})(\hat{x}^s_{l|l-1} - \hat{x}_{l|l-1})^\intercal] \\ =& P + \mathbb{E}[(\hat{x}^s_{l|l-1} - \hat{x}_{l|l-1})(\hat{x}^s_{l|l-1} - \hat{x}_{l|l-1})^\intercal], \end{aligned} \tag{17}$$

where the last equality holds due to the orthogonality principle, i.e., all the random variables generated by the smart sensor are independent of the estimation error $x_l - \hat{x}^s_{l|l-1}$ of the MMSE estimate $\hat{x}^s_{l|l-1}$ (Bai, Gupta, et al., 2017). More specifically, $\hat{x}^s$ is the state estimate of the smart sensor. $\hat{x}$ is the state estimate of the remote estimator and it is updated by the modified innovation $\tilde{z}_k$, where $\tilde{z}_k$ is linear with the innovation of z_k. Since the error vector $x_l - \hat{x}^s_{l|l-1}$ is orthogonal to the innovation z_k, the last equality holds. Define $\tilde{e}_l \triangleq \hat{x}^s_{l|l-1} - \hat{x}_{l|l-1}$, where

$$\begin{aligned} \hat{x}^s_{l|l-1} &= A\hat{x}^s_{l-1|l-2} + AKz_{l-1} \\ &= A^{l-1}\hat{x}_{1|0} + \sum_{i=1}^{l-1} A^i K z_{l-i}, \end{aligned}$$

and

$$\begin{aligned} \hat{x}_{l|l-1} &= A\hat{x}_{l-1|l-2} + AK\tilde{z}_{l-1} \\ &= A^{l-1}\hat{x}_{1|0} + \sum_{i=1}^{l-1} A^i K \tilde{z}_{l-i}. \end{aligned} \tag{18}$$

Since $\hat{x}^s_{1|0} = \hat{x}_{1|0}$, which implies that $\tilde{e}_1 = \mathbf{0}_{m\times 1}$, we have

$$\mathbb{E}[\tilde{e}_l \tilde{e}_l^\intercal] = \sum_{i=1}^{l-1} A^i K \mathbb{E}\left[\left(z_{l-i} - \tilde{z}_{l-i}\right)\left(z_{l-i} - \tilde{z}_{l-i}\right)^\intercal\right](A^i K)^\intercal. \tag{19}$$

Take the limit of $\mathbb{E}[\tilde{e}_l \tilde{e}_l^\intercal]$, we have

$$\begin{aligned} &\lim_{l\to\infty} \mathbb{E}[\tilde{e}_l \tilde{e}_l^\intercal] \\ &= \lim_{l\to\infty} \sum_{i=1}^{l-1} A^i K \mathbb{E}\left[\left[(I_m - S)z_{l-i} - \phi_{l-i}\right]\left[(I_m - S)z_{l-i} - \phi_{l-i}\right]^\intercal\right](A^i K)^\intercal \\ &= \lim_{l\to\infty} \sum_{i=1}^{l-1} A^i K \left[(I_m - S)\Sigma_z (I_m - S)^\intercal + \Sigma_z - S\Sigma_z S^\intercal\right] K^\intercal (A^i)^\intercal. \end{aligned} \tag{20}$$

For the simplicity of notations, we define

$$\mathcal{P}_1 \triangleq K\left[(I_m - S)\Sigma_z(I_m - S)^\intercal + \Sigma_z - S\Sigma_z S^\intercal\right]K^\intercal.$$

Since $\mathcal{P}_1$ is positive semi-definite and A is stable, (21) can be simplified as

$$\lim_{l\to\infty}\mathbb{E}[\tilde{e}_l\tilde{e}_l^\intercal] = \mathcal{Y},$$

where $\mathcal{Y} = \mathcal{Y}_1 - \mathcal{P}_1$ and $\mathcal{Y}_1$ is the solution to the discrete Lyapunov equation $\mathcal{Y}_1 = A\mathcal{Y}_1A^\intercal + \mathcal{P}_1$.

The performance degradation ratio is

$$\eta_s = \lim_{k\to\infty}\frac{\frac{1}{k}\sum_{l=1}^{k}\operatorname{tr}\tilde{P}_l}{\operatorname{tr}P} = \frac{\operatorname{tr}P + \operatorname{tr}\mathcal{Y}}{\operatorname{tr}P}.$$

Considering $\Sigma_z - S\Sigma_z S^\intercal \succeq$ and the expression of $\mathcal{P}_1$, we can derive two special cases as follows:

(1) $(T, S, \Phi) = (\mathbf{0}_{m\times m}, I_m, \mathbf{0}_{m\times m})$: $\tilde{z}_l = z_l$, i.e., the attacker is not launching an attack, the corresponding ratio $\eta_{s,1} = 1$.
(2) $(T, S, \Phi) = (\mathbf{0}_{m\times m}, -I_m, \mathbf{0}_{m\times m})$: $\tilde{z}_l = -z_l$, i.e., the attacker flips the sign of the innovation.

For case (2), we can easily derive that

$$\eta_{s,2} = \lim_{k\to\infty}\frac{\frac{1}{k}\sum_{l=1}^{k}\operatorname{tr}\tilde{P}_l}{\operatorname{tr}P} = 1 + 4\frac{\operatorname{tr}\mathcal{X}}{\operatorname{tr}P} > 1.$$

where $\mathcal{X} = \mathcal{X}_1 - \mathcal{P}_2$, $\mathcal{X}_1$ is the solution to $\mathcal{X}_1 = A\mathcal{X}_1A^\intercal + \mathcal{P}_2$ and $\mathcal{P}_2$ is defined as $\mathcal{P}_2 \triangleq K\Sigma_z K^\intercal$. Let us compare $\eta_{s,2}$ and η_s.

$$\begin{aligned}
&4\operatorname{tr}\mathcal{X} - \operatorname{tr}\mathcal{Y}\\
&= \sum_{i=1}^{\infty}\operatorname{tr}\left(A^iK\left(2\Sigma_z + S\Sigma_z + \Sigma_z S^\intercal\right)(A^iK)^\intercal\right)\\
&= \sum_{i=1}^{\infty}\operatorname{tr}\left(A^iK\left((I_m + S)\Sigma_z(I_m + S)^\intercal - S\Sigma_z S^\intercal + \Sigma_z\right)(A^iK)^\intercal\right) \geq 0.
\end{aligned}$$

The worst performance degradation ratio is $\eta_s = 1 + 4\frac{\operatorname{tr}\mathcal{X}}{\operatorname{tr}P}$ with the corresponding attack strategy $(T^*, S^*, \Phi^*) = (\mathbf{0}_{m\times m}, -I_m, \mathbf{0}_{m\times m})$. □

Remark 3.1. In Theorem 3.1, the linear first-order attack model (5) is considered. The same result can be easily extended to a general linear time-invariant attack model of the form:

$$\begin{aligned}
c_k &= Mc_{k-1} + Nz_{k-1},\\
\tilde{z}_k &= Wc_k + Gz_k,
\end{aligned}$$

where $c_k \in \mathbb{R}^m$, $M \in \mathbb{R}^{m\times m}$, $N \in \mathbb{R}^{m\times p}$, $W \in \mathbb{R}^{p\times m}$, $G \in \mathbb{R}^{p\times p}$. That is, under strictly stealthy attacks of the above form, the worst case performance remains the same as that in Theorem 3.1 and the optimal attack strategy is that $G = -I_p$ and the parameters M, N and W need to satisfy $WM^iN = 0$ $(i = 0, 1, \dots)$.

Remark 3.2. For scalar systems, the worst performance degradation ratio is

$$\eta_s = 1 + \frac{4A^2K^2(C^2P + R)}{(1 - A^2)P}$$

and the corresponding attack strategy is $(T^*, S^*, \Phi^*) = (0, -1, 0)$. Hence, the degradation is worse for systems with a higher Kalman filter gain. Note also that the worst case attack simply flips the sign of the innovation sequence.

Remark 3.3. Under the strict stealthiness metric, the optimal attack strategy in our work is aligned with the result about the worst-case linear attack under the χ^2 false alarm detector obtained in Guo et al. (2016). The reason why the optimal attack policies are the same for the different problem settings is that the modified innovation needs to preserve the statistics of the attack-free innovation, which leads to that $T = \mathbf{0}_{m\times m}$. However, since we consider a more general model that utilizes both past and current information, the derivation of the optimal attack strategy is different from that in Guo et al. (2016). Note that $\eta = 1$ when $A = \mathbf{0}_{n\times n}$.

Remark 3.4. Under Assumption 2.1, i.e., A is stable, Theorem 3.1 provides a closed-form solution for the performance degradation ratio. If A is not stable, (21) will diverge. Besides, although we mainly study the scenario under strictly stealthy attacks in this section, the strictly stealthy attack can be considered as a special case of ϵ-stealthy attacks with $\epsilon = 0$. In other words, a strictly stealthy attack strategy should be feasible when considering an ϵ-stealthy attack.

4. ϵ-Stealthy attacks

In this section, we will characterize the maximal performance degradation under an ϵ-stealthy attack. The memoryless attacker $T = \mathbf{0}_{m\times m}$ was studied in Guo et al. (2018), Li and Yang (2019). We focus on the attacker with memory, i.e., $T \neq \mathbf{0}_{m\times m}$. For the sake of analysis, we will focus on scalar systems, i.e., $m = n = 1$ in the following analysis. The vector case will be a potential future work. In order to differentiate between scalar and vector systems, we use σ_z^2 to replace Σ_z to represent the covariance of z_k, i.e., $\sigma_z^2 = C^2P + R$.

For the simplicity of notation, define

$$q \triangleq \frac{\Phi}{\sigma_z^2}, \qquad q \geq 0.$$

Then we have the following lemma, the proof of which is reported in the appendix.

Lemma 4.1. *Consider the scalar system (1)–(2), the optimization problem (15) is equivalent to the following problem:*

$$\begin{aligned}
&\arg\max_{T,S,q} \quad J(T, S, q),\\
&\text{s. t.} \quad -\frac{1}{2} - \frac{1}{2}\log(S^2 + q) + \frac{S^2 + q}{2(1 - T^2)} = \epsilon,\\
&\qquad -S_{oq\max} < S \leq -\sqrt{e^{-2\epsilon} - q},
\end{aligned} \tag{21}$$

where

$$J(T, S, q) = (1 - S)^2 + q + \frac{T^2(S^2 + q)}{1 - T^2} - 2AT\frac{S - S^2 - ST^2 - q}{(1 - T^2)(1 - AT)}.$$

For a given q, denote

$$J_{q1}(T, S) \triangleq J(T, S, q). \tag{22}$$

From the constraint function of (21), one can obtain

$$T = f_q(S) \triangleq \sqrt{1 - \frac{S^2 + q}{2\epsilon + 1 + \log(S^2 + q)}}. \tag{23}$$

By substituting (23) into (22), we have

$$\begin{aligned}
J_{q1}(f_q(S), S) = J_{q2}(S) \triangleq &-(2\epsilon + \log(S^2 + q)) - \frac{2S}{1 - Af_q(S)}\\
&+ \frac{2(2\epsilon + 1 + \log(S^2 + q))}{1 - Af_q(S)}.
\end{aligned} \tag{24}$$

The following lemma characterizes the worst performance ratio for the estimation error covariance and gives the corresponding attack strategy to achieve this performance bound for a given q, the proof of which is reported in the appendix.

Lemma 4.2. *Consider scalar system (1)–(2) satisfying Assumption 2.1 and linear attack of the form (5). Given $q \geq 0$ and $\epsilon > 0$, under the ϵ-stealthy attacks, the worst performance degradation ratio for the estimation error covariance is*

$$\eta_\epsilon = 1 + \frac{J_{q\,\mathrm{opt}} A^2 K^2 \sigma_z^2}{(1-A^2)P},$$

where $J_{q\,\mathrm{opt}} = J_{q2}(S_q)$ with S_q being such that $J'_{q2}(S_q) = 0$. The corresponding attack strategy is (T_q, S_q), where $T_q = f_q(S_q)$.

Next, we will first prove that the optimal strategy requires $q = 0$. Then, we provide the optimal attack strategy and the corresponding worst case performance. Finally, we show that our proposed attack strategy can achieve a better attack performance than that of the existing work in Guo et al. (2018) under the same ϵ-stealthy attacks. Similarly, for the sake of readability, the proof of the following lemma is reported in the appendix.

Lemma 4.3. *The solution to the original optimization problem (15) requires $q = 0$. Hence, the optimization problem (21) can be transformed into the following problem:*

$$\begin{aligned} \arg\max_{S,T} \quad & J(S,T,0), \\ \text{s. t.} \quad & -\frac{1}{2} - \frac{1}{2}\log(S^2) + \frac{S^2}{2(1-T^2)} = \epsilon, \\ & 0 < |T| \leq \sqrt{1-e^{-2\epsilon}}. \end{aligned}$$

Before we give the theorem regarding ϵ-stealthy attacks, we define the following equations for the simplicity of notations:

$$f_0(S) \triangleq \sqrt{1 - \frac{S^2}{2\epsilon + 1 + \log(S^2)}},$$

$$J_0(S) \triangleq -(2\epsilon + \log(S^2)) - \frac{2S}{1 - A f_0(S)} + \frac{2(2\epsilon + 1 + \log(S^2))}{1 - A f_0(S)}.$$

The following theorem characterizes the maximal performance degradation ratio under an ϵ-stealthy attack. We also provide the attack strategy to achieve the maximum.

Theorem 4.1. *Consider the scalar system (1)–(2) satisfying Assumption 2.1 and linear attack of the form (5). Given $\epsilon > 0$, under the ϵ-stealthy attacks, the worst performance degradation ratio for the estimation error covariance is*

$$\eta_\epsilon = 1 + \frac{J_{\mathrm{opt}} A^2 K^2 \sigma_z^2}{(1-A^2)P},$$

where $J_{\mathrm{opt}} = J_0(S_{\mathrm{opt}})$. The corresponding attack strategy is $(T_{\mathrm{opt}}, S_{\mathrm{opt}}, 0)$, where S_{opt} satisfies $J'_0(S_{\mathrm{opt}}) = 0$ and $T_{\mathrm{opt}} = f_0(S_{\mathrm{opt}})$.

Proof. The results follow from Lemmas 4.2 and 4.3.

Remark 4.1. The attack policy in Guo et al. (2018) is given by $\tilde{z}_k = \sqrt{X} z_k$, where X is the largest solution of the equation $X = 2\epsilon + 1 + \log X$. It corresponds for our model to $q_g = 0, T_g = 0, S_g = -\sqrt{X}$. The corresponding performance degradation ratio is

$$\eta_{\epsilon,g} = 1 + \frac{(1-S_g)^2 A^2 K^2 \sigma_z^2}{(1-A^2)P}.$$

Hence, the difference of performance degradation between our approach and that in Guo et al. (2018) is given by:

$$\eta_s - \eta_{\epsilon,g} = \frac{(J_{\mathrm{opt}} - (1-S_g)^2) A^2 K^2 \sigma_z^2}{(1-A^2)P},$$

where J_{opt} is defined in Theorem 4.1. Note that the optimal parameter S for our proposed approach is between $-S_{oq\,\max}$ and $-e^{-\epsilon}$ while the existing linear attack strategy takes $-S_{oq\,\max}$, where $S_{oq\,\max}$ is defined in the proof of Lemma 4.2 and $J_{opt} \geq (1 - S_g)^2$. Hence, it is clear that the performance degradation ratio bound for the estimation error covariance induced by the proposed attack strategy is larger than or equal to the existing linear attack strategy in Guo et al. (2018) under ϵ-stealthy attacks with the same ϵ.

Remark 4.2. We focus on the scalar case in this section. For the vector case, Lemma A.2 needs to be rewritten as "If an attacker employs an ϵ-stealthy attack in the form of (5), then $\rho(T) < 1$". In Lemma A.3, the derivation of the objective function involves the sum of a geometric sequence. For the vector case, we need to use Proposition 1.5.31 in Hubbard and Hubbard (2015), "Let A be a square matrix. If $\rho(A) < 1$, the series $S = I + A + A^2 + \cdots$ converges to $(I - A)^{-1}$". Reconsider Lemma A.3, since $\rho(A) < 1$ and $\rho(T) < 1$, the above proposition can be directly applied. Then, the optimization problem for the vector case can be obtained by using a similar method. However, it is difficult to obtain a closed-form solution since the optimization problem is not convex and involves more than one parameter. Further studies will be carried out in the future.

5. Simulation

In this section, we provide some numerical examples to evaluate the performance of the proposed attack strategies.

5.1. Vector case under strictly stealthy attacks

In this subsection, we set the system parameters as follows:

$$A = \begin{bmatrix} 0.5 & 0.2 \\ 0.1 & 0.8 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix},$$

$$Q = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.3 \end{bmatrix}, \quad R = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.6 \end{bmatrix}.$$

We can obtain that

$$K = \begin{bmatrix} 0.3583 & -0.2866 \\ 0.2374 & 0.2027 \end{bmatrix}, \quad P = \begin{bmatrix} 0.6833 & -0.0302 \\ -0.0302 & 0.3548 \end{bmatrix},$$

and from Theorem 3.1, the optimal attack performance degradation ratio is $\eta = 4.6017$. Assume that the attack starts at time $k = 53$. We run 10000 simulations. The ratio of the state estimation error covariance $\tilde{P}$ to P v.s. time k is shown in Fig. 2. The parameter S_i $(i = 1, 2)$ for attack 1–2 and S^* for strictly stealthy attack and S_{normal} for normal operation are as follows:

$$S^* = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, S_{\mathrm{normal}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

$$S_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}, S_2 = \begin{bmatrix} -0.5 & 0 \\ 0.1 & -0.8 \end{bmatrix},$$

and the corresponding covariance of the added noise is derived by $\Phi = \Sigma_z - S\Sigma_z S^\intercal$.

From this figure, there is an obvious difference of the performance degradation between the normal operation and an attack operation. It is easy to see that the error covariance ratio under the optimal attack is larger than the one under normal operation and other attacks with different attack parameters. Besides, we can also see that the optimal simulation value is almost the same as the theoretical value.

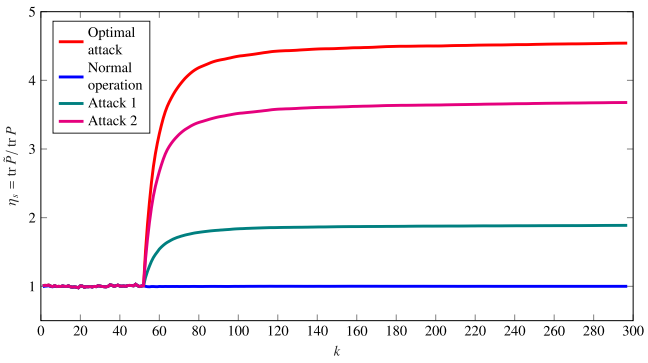


Fig. 2. The ratio of the error covariance $\tilde{P}$ to P v.s. time k. The red line is the ratio of simulation under strictly stealthy attack. The blue line is the ratio of the simulation under normal operation. The teal and magenta lines denote the corresponding ratio under different attack type 1 to attack type 2, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

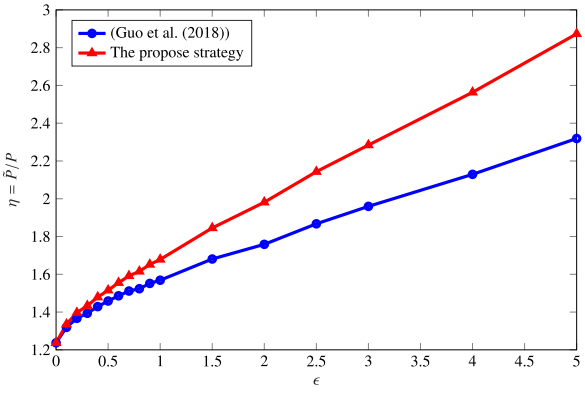


Fig. 3. The ratio of the error covariance $\tilde{P}$ to P v.s. stealthiness level ϵ. The blue line with circle markers is the ratio obtained from the existing work (Guo et al., 2018). The red line with upward-pointing triangle markers denotes the ratio in our work. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

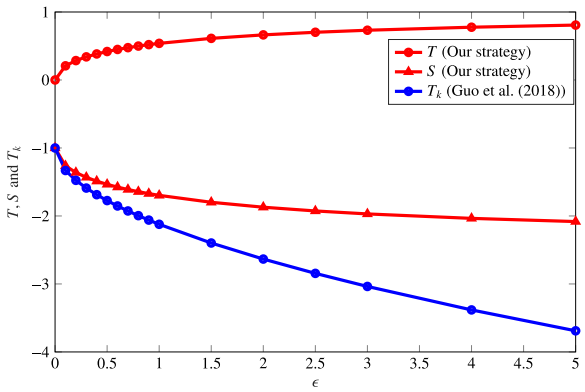


Fig. 4. The values of T, S and T_k (which is used in Guo et al. (2018)) v.s. the stealthiness level ϵ. The red lines with circle markers and triangle markers are the value of T and S in our proposed strategy, respectively. The blue line with circle markers denotes the value of T_k from the existing work (Guo et al., 2018). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

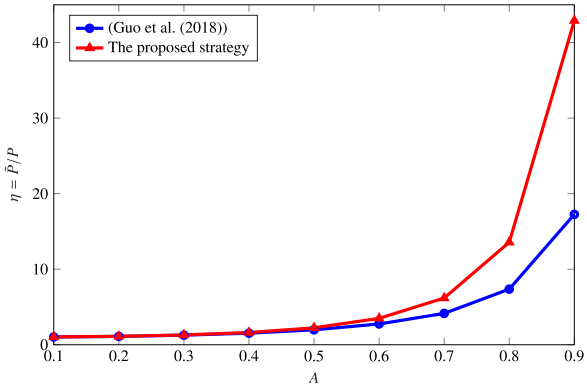


Fig. 5. The ratio of the error covariance $\tilde{P}$ to P v.s. A. The blue line with circle markers is the ratio obtained in the existing work (Guo et al., 2018). The red line with upward-pointing triangle markers denotes the ratio in our work. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.2. Different ϵ-stealthy level

In this subsection, we consider an LTI system with scalars and set $A = 0.4$, $C = 1$, $Q = 0.2$, and $R = 0.5$. One can compute that $K = 0.3102$, and $P = 0.2248$.

The ratio of the state estimation error covariance $\tilde{P}$ to P v.s. stealthiness level ϵ is shown in Fig. 3. From this figure, one could see that the error covariance obtained in our work is equal to the one obtained in the existing work (Guo et al., 2018) when $\epsilon = 0$. And the error covariance obtained in our work is larger than the one derived in Guo et al. (2018) when $\epsilon > 0$. Furthermore, the difference of the error covariances between our work and (Guo et al., 2018) is becoming larger as ϵ grows.

The values of T, S and T_k (which is used in Guo et al., 2018) v.s. the stealthiness level ϵ are shown in Fig. 4. From Fig. 4, one can see that as ϵ grows, the absolute values of T and S are becoming larger. It means that as the stealthiness level ϵ increases, the attacker employs more past attack information and current innovation.

5.3. Different system parameter A

In this subsection, we consider an LTI system with scalars and set $\epsilon = 0.8$, $C = 1$, $Q = 0.2$, and $R = 0.5$. We study the difference induced by different system parameter A. The ratio of the error covariance $\tilde{P}$ to P v.s. A is shown in Fig. 5. From this figure, one could see that the error covariance obtained in our work is larger than the one derived in Guo et al. (2018). Furthermore, the difference of the error covariances between our work and (Guo et al., 2018) is becoming larger with A increasing.

The values of T, S and T_k (which is used in Guo et al., 2018) v.s. the system matrix A are shown in Fig. 6. From this figure, one can see that as A increases, the absolute value of T is becoming larger and the absolute value of S is becoming smaller. It implies that as the system parameter A increases, the remote state estimator will attach more importance to the priori state estimate by (6) and (7). Correspondingly, the attacker will employ the past information more and use current innovation less in order to maximize the attack performance. Since the proposed approach in Guo et al. (2018) is only related with the stealthiness level, the value of T_k keeps constant.

6. Conclusion

In this paper, we characterized the fundamental limits for innovation-based remote state estimation under linear attacks. The attacker was constrained to follow a linear attack type based on the past attack signal, the latest innovation and an additive random variable. We obtained optimal attack strategies to

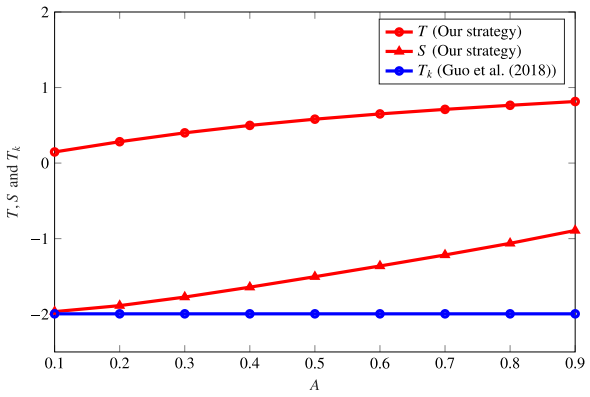


Fig. 6. The values of T, S and T_k (which is used in Guo et al., 2018) v.s. A. The red lines with circle markers and triangle markers are the value of T and S in our proposed strategy, respectively. The blue line with circle markers denotes the value of T_k from the existing work (Guo et al., 2018). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

achieve maximal performance degradation under a given stealthiness requirement. Then we provided the maximal performance degradation ratio and the corresponding optimal attack strategy to achieve this maximum under strictly stealthy attacks for vector systems, which is a generalization of the previous work. For ϵ-stealthy attacks on scalar systems, the optimal attack strategy with an additive random noise was also presented. It was proven that the maximal performance degradation ratio can be achieved without additive noise and the proposed strategy performs better than the existing linear attack strategies in terms of performance degradation. Simulation results were presented to support the theoretical results. For future works, we would like to study vector systems under ϵ-stealthy attacks. Besides, it is also of great interest to study multi-sensor extensions, and in such extensions investigate how sensors can collaborate to mitigate attacks under various adversarial scenarios.

Acknowledgment

The authors are grateful to Yuchao Li from KTH Royal Institute of Technology for the insightful comments.

Appendix A. Proof of Lemma 4.1

The whole section is devoted to proving Lemma 4.1. We shall present several lemmas and then proceed with the proof of Lemma 4.1.

First, we give the following lemma to characterize the property of the modified innovation sequence, which will be used to simplify the constraint condition of the optimization problem (15). The following lemma is for a vector case, and the scalar case follows as a special case.

Lemma A.1. *If an attacker employs an attack in the form of (5), the differential entropy of the compromised innovation sequence $\tilde{z}_1^k$ is equal to $\frac{k}{2}\log\left((2\pi e)^p \det \mathcal{S}\right)$, where $\mathcal{S} \triangleq S\Sigma_z S^\intercal + \Phi$.*

Proof. Here, we use the notation $h(\tilde{z}_1^k)$ to represent the differential entropy:

$$h(\tilde{z}_1^k) = -\int f_{\tilde{z}_1^k}(\xi)\log f_{\tilde{z}_1^k}(\xi)d\xi,$$

where $f_{\tilde{z}_1^k}$ is the probability density function.

By (16), $\tilde{z}_1^k$ follows a multivariate Gaussian distribution. We have:

$$h(\tilde{z}_1^k) = \frac{1}{2}\log\left((2\pi e)^{pk}\det\Sigma\right), \tag{25}$$

where

$$\Sigma \triangleq \mathrm{Cov}\left([\tilde{z}_1^\intercal, \tilde{z}_2^\intercal, \ldots, \tilde{z}_k^\intercal]^\intercal\right) = \begin{bmatrix} \mathcal{S} & \mathcal{S}T^\intercal & \cdots & \mathcal{S}(T^{k-1})^\intercal \\ T\mathcal{S} & T\mathcal{S}T^\intercal + \mathcal{S} & \cdots & T\mathcal{S}(T^{k-1})^\intercal + \mathcal{S}(T^{k-2})^\intercal \\ \vdots & \vdots & \ddots & \vdots \\ T^{k-1}\mathcal{S} & T^{k-1}\mathcal{S}T^\intercal + T^{k-2}\mathcal{S} & \cdots & T^{k-1}\mathcal{S}(T^{k-1})^\intercal + \cdots + \mathcal{S} \end{bmatrix}, \tag{26}$$

and $\mathcal{S} \triangleq S\Sigma_z S^\intercal + \Phi$. One can perform an elementary row transformation on the matrix Σ and obtain $\det\Sigma = (\det\mathcal{S})^k$.

Hence, for any T, the differential entropy can be obtained as follows:

$$h(\tilde{z}_1^k) = \frac{k}{2}\log\left((2\pi e)^p \det\mathcal{S}\right). \tag{27}$$

The proof is completed. □

Lemma A.2. *If an attacker employs an ϵ-stealthy attack in the form of (5), then $|T| < 1$.*

Proof. From Lemma A.1, it is easy to obtain

$$\begin{aligned}
&\frac{1}{k}D(\tilde{z}_1^k \parallel z_1^k) \\
&= -\frac{1}{k}h(\tilde{z}_1^k) + \frac{1}{2}\log(2\pi\sigma_z^2) + \frac{1}{k}\sum_{l=1}^{k}\frac{E[(\tilde{z}_l)^2]}{2\sigma_z^2} \\
&= -\frac{1}{2}\log\left(2\pi e(S^2\sigma_z^2 + \Phi)\right) + \frac{1}{2}\log(2\pi\sigma_z^2) + \frac{1}{k}\sum_{l=1}^{k}\frac{E[(\tilde{z}_l)^2]}{2\sigma_z^2} \\
&= -\frac{1}{2} - \frac{1}{2}\log\left(\frac{S^2\sigma_z^2 + \Phi}{\sigma_z^2}\right) + \frac{1}{k}\sum_{l=1}^{k}\frac{E[(\tilde{z}_l)^2]}{2\sigma_z^2}.
\end{aligned}$$

Let us consider the sufficient condition of ϵ-stealthy:

$$\lim_{k\to\infty}\frac{1}{k}D(\tilde{z}_1^k \parallel z_1^k) \le \epsilon, \tag{28}$$

which implies

$$\lim_{k\to\infty} -\frac{1}{2} - \frac{1}{2}\log\left(\frac{S^2\sigma_z^2 + \Phi}{\sigma_z^2}\right) + \frac{1}{k}\sum_{l=1}^{k}\frac{E[(\tilde{z}_l)^2]}{2\sigma_z^2} \le \epsilon,$$

where $\mathbb{E}[(\tilde{z}_l)^2] = \sum_{i=0}^{l-1} T^{2i}\left(S^2\sigma_z^2 + \Phi\right)$.

Similarly, we divide four cases ($0 < |T| < 1$, $|T| = 1$ and $|T| > 1$) to compute $\mathbb{E}[(\tilde{z}_l)^2]$:

1. $0 < |T| < 1$:

$$\mathbb{E}[(\tilde{z}_l)^2] = \frac{1 - T^{2l}}{1 - T^2}\left(S^2\sigma_z^2 + \Phi\right), \tag{29}$$

then we have $\lim_{k\to\infty}\frac{1}{k}\sum_{l=1}^{k}\frac{E[(\tilde{z}_l)^2]}{2\sigma_z^2} = \frac{S^2\sigma_z^2+\Phi}{2(1-T^2)\sigma_z^2}$, which could satisfy the requirement of ϵ-stealthiness.
2. $|T| = 1$: $\mathbb{E}[(\tilde{z}_l)^2] = l\left(S^2\sigma_z^2 + \Phi\right)$, then, $\lim_{k\to\infty}\frac{1}{k}\sum_{l=1}^{k}\frac{\mathbb{E}[(\tilde{z}_l)^2]}{2\sigma_z^2} \to \infty$, which contradicts the requirement of ϵ-stealthiness.
3. $|T| > 1$: the sum is expressed as (29). It is easy to check that $\lim_{k\to\infty}\frac{1}{k}\sum_{l=1}^{k}\frac{E[(\tilde{z}_l)^2]}{2\sigma_z^2}$ will diverge, which also contradicts the requirement of ϵ-stealthiness.

As a result, T must satisfy that $0 < |T| < 1$. □

Lemma A.3. *The optimization problem (15) is equivalent to the following problem:*

$$\begin{aligned}\arg\max_{T,S,q}\quad & (1-S)^2+q+\frac{T^2(S^2+q)}{1-T^2}-2AT\frac{S-S^2-ST^2-q}{(1-T^2)(1-AT)},\\ \text{s.t.}\quad & -\frac{1}{2}-\frac{1}{2}\log(S^2+q)+\frac{S^2+q}{2(1-T^2)}\le\epsilon,\\ & 0<|T|<1.\end{aligned}\tag{30}$$

Proof. From (17), one can see that the error covariance between the state estimate and the real state, $\tilde{P}_l$, can be split into two parts, one is the minimum mean square error P which is constant, and the other is the error covariance of $\tilde{e}_{l+1}=\tilde{x}^s_{l+1|l}-\tilde{x}_{l+1|l}$. Note that

$$\begin{aligned}\tilde{e}_{k+1}=&\hat{x}^s_{k+1|k}-\hat{x}_{k+1|k}\\ =&A\hat{x}^s_{k|k-1}+AKz_k-(A\hat{x}_{k|k-1}+AK\tilde{z}_k)\\ =&A(\hat{x}^s_{k|k-1}-\hat{x}_{k|k-1})-AK(T\tilde{z}_{k-1}+Sz_k+\phi_k)+AKz_k\\ =&A\tilde{e}_k+AK(1-S)z_k-AKT\tilde{z}_{k-1}-AK\phi_k.\end{aligned}\tag{31}$$

From (31), one can know that $\mathbb{E}[\tilde{e}_k]=0$. Hence, the covariance of $\tilde{e}_k$ is

$$\begin{aligned}&\mathbb{E}[(\tilde{e}_{k+1})^2]\\ =&A^2\mathbb{E}[(\tilde{e}_k)^2]+[AK(1-S)]^2\sigma_z^2+2A^2K(1-S)\mathbb{E}[\tilde{e}_kz_k]\\ &+(AKT)^2\mathbb{E}[(\tilde{z}_{k-1})^2]+A^2K^2q\sigma_z^2-2A^2KT\mathbb{E}[\tilde{e}_k\tilde{z}_{k-1}]\\ \overset{(a)}{=}&A^2\mathbb{E}[(\tilde{e}_k)^2]+[AK(1-S)]^2\sigma_z^2+(AKT)^2\mathbb{E}[(\tilde{z}_{k-1})^2]\\ &+A^2K^2q\sigma_z^2-2A^2KT\mathbb{E}[\tilde{e}_k\tilde{z}_{k-1}]\\ =&A^2\mathbb{E}[(\tilde{e}_k)^2]+[AK(1-S)]^2\sigma_z^2-2A^2KT\mathbb{E}[\tilde{e}_k\tilde{z}_{k-1}]\\ &+A^2K^2q\sigma_z^2+(AKT)^2\frac{1-T^{2(k-1)}}{1-T^2}(S^2+q)\sigma_z^2,\end{aligned}\tag{32}$$

where

$$\begin{aligned}\tilde{e}_k=&A\tilde{e}_{k-1}+AK(1-S)z_{k-1}-AKT\tilde{z}_{k-2}-AK\phi_{k-1}\\ =&A^{k-1}\tilde{e}_1+AK\left(\sum_{i=1}^{k-1}A^{k-1-i}(1-S)z_i\right)\\ &-AK\left(\sum_{i=0}^{k-2}A^{k-2-i}T\tilde{z}_i\right)-AK\left(\sum_{i=1}^{k-1}A^{k-1-i}\phi_i\right),\end{aligned}$$

and (a) holds due to the independence of $\tilde{e}_k$ and z_k.

To simplify the notations, we define

$$\begin{aligned}\Xi_1\triangleq&AK(1-S)\mathbb{E}\left[\left(\sum_{i=1}^{k-1}A^{k-1-i}z_i\right)\tilde{z}_{k-1}\right],\\ \Xi_2\triangleq&AK\mathbb{E}\left[\left(\sum_{i=0}^{k-2}A^{k-2-i}T\tilde{z}_i\right)\tilde{z}_{k-1}\right],\\ \Xi_3\triangleq&AK\mathbb{E}\left[\left(\sum_{i=1}^{k-1}A^{k-1-i}\phi_i\right)\tilde{z}_{k-1}\right],\end{aligned}$$

then we have

$$\begin{aligned}\Xi_1=&\frac{1-(AT)^{k-1}}{1-AT}AK(1-S)S\sigma_z^2,\\ \Xi_2=&\frac{AT\left[1-(AT)^{k-2}\right]}{1-AT}\frac{KT(S^2+q)\sigma_z^2}{1-T^2}\\ &-\frac{AT^k(T^{k-2}-A^{k-2})}{T-A}\frac{KT(S^2+q)\sigma_z^2}{1-T^2}\end{aligned}$$

$$\Xi_3=\frac{1-(AT)^{k-1}}{1-AT}AKq\sigma_z^2,$$

Reconsider the third term of (32), we have

$$\begin{aligned}&\mathbb{E}[\tilde{e}_k\tilde{z}_{k-1}]\\ =&\Xi_1-\Xi_2-\Xi_3\\ =&\frac{1-(AT)^{k-1}}{1-AT}AK(1-S)S\sigma_z^2-\frac{1-(AT)^{k-1}}{1-AT}AKq\sigma_z^2\\ &+\frac{AT^k(T^{k-2}-A^{k-2})}{T-A}\frac{KT(S^2+q)\sigma_z^2}{1-T^2}\\ &-\frac{AT\left[1-(AT)^{k-2}\right]}{1-AT}\frac{KT(S^2+q)\sigma_z^2}{1-T^2}.\end{aligned}$$

Consider the asymptotic behavior for (32) and take the limit for the above equation, one can obtain

$$\begin{aligned}&\lim_{k\to\infty}\frac{1}{k}\sum_{l=1}^{k}\mathbb{E}[(\tilde{e}_{l+1})^2]\\ =&\lim_{k\to\infty}\frac{1}{k}\sum_{l=1}^{k}A^2\mathbb{E}[(\tilde{e}_n)^2]+A^2K^2(1-S)^2\sigma_z^2+A^2K^2q\sigma_z^2\\ &+(AKT)^2\frac{(S^2+q)\sigma_z^2}{1-T^2}-2A^2KT\frac{1}{k}\sum_{l=1}^{k}\mathbb{E}[\tilde{e}_n\tilde{z}_{l-1}]\\ =&\lim_{k\to\infty}\frac{1}{k}\sum_{l=1}^{k}A^2\mathbb{E}[(\tilde{e}_n)^2]+A^2K^2\left[(1-S)^2+q\right.\\ &\left.+T^2\frac{(S^2+q)}{1-T^2}\right]\sigma_z^2-2A^2KT\frac{1}{k}\sum_{l=1}^{k}\mathbb{E}[\tilde{e}_n\tilde{z}_{l-1}]\\ =&\lim_{k\to\infty}\frac{1}{k}\sum_{l=1}^{k}A^2\mathbb{E}[(\tilde{e}_n)^2]+A^2K^2\left[(1-S)^2+q+T^2\frac{(S^2+q)}{1-T^2}\right]\sigma_z^2\\ &-2A^2KT\left[\frac{AK(1-S)S}{1-AT}-\frac{AKT^2(S^2+q)}{(1-T^2)(1-AT)}-\frac{AKq}{1-AT}\right]\sigma_z^2.\end{aligned}$$

From (32), it is easy to obtain (33) since $\lim_{k\to\infty}\frac{1}{k}\mathbb{E}[(\tilde{e}_1)^2]=0$ and $\lim_{k\to\infty}\frac{1}{k}\mathbb{E}[(\tilde{e}_{k+1})^2]=0$.

$$\begin{aligned}&\lim_{k\to\infty}\frac{1-A^2}{k}\sum_{l=1}^{k}E[(\tilde{e}_{l+1})^2]\\ =&\lim_{k\to\infty}\frac{A^2}{k}E\left[(\tilde{e}_1)^2-(\tilde{e}_{k+1})^2\right]\\ &+A^2K^2\left[(1-S)^2+q+T^2\frac{(S^2+q)}{1-T^2}\right]\sigma_z^2\\ &-2A^2KT\left[\frac{AK(1-S)S}{1-AT}-\frac{AKT^2(S^2+q)}{(1-T^2)(1-AT)}-\frac{AKq}{1-AT}\right]\sigma_z^2\\ =&A^2K^2\left[(1-S)^2+q+T^2\frac{(S^2+q)}{1-T^2}\right]\sigma_z^2\\ &-2A^2KT\left[\frac{AK(1-S)S}{1-AT}-\frac{AKT^2(S^2+q)}{(1-T^2)(1-AT)}-\frac{AKq}{1-AT}\right]\sigma_z^2.\end{aligned}\tag{33}$$

Hence, the optimization problem can be rewritten as

$$\begin{aligned}\arg\max_{T,S,q}\quad & \sigma_z^2A^2K^2\left[\left[(1-S)^2+q+T^2\frac{(S^2+q)}{1-T^2}\right]\right.\\ &\left.-2AT\left[\frac{(1-S)S}{1-AT}-\frac{T^2(S^2+q)}{(1-T^2)(1-AT)}-\frac{q}{1-AT}\right]\right],\end{aligned}$$

$$\text{s.t.}\quad -\frac{1}{2}-\frac{1}{2}\log(S^2+q)+\frac{S^2+q}{2(1-T^2)}\le\epsilon,$$
$$0<|T|<1.$$

Since $\sigma_z^2>0$ and $A^2K^2>0$, the optimization problem can be simplified as follows:

$$\arg\max_{T,S,q}\ (1-S)^2+q+\frac{T^2(S^2+q)}{1-T^2}-\frac{2AT(S-ST^2-S^2-q)}{(1-T^2)(1-AT)},$$
$$\text{s.t.}\quad -\frac{1}{2}-\frac{1}{2}\log(S^2+q)+\frac{S^2+q}{2(1-T^2)}\le\epsilon,$$
$$0<|T|<1.$$

The proof is completed. □

Lemma A.4. *When S is negative, q is fixed and the absolute value of T is fixed, $J(T,S,q)\ge J(-T,S,q)$, where the sign of T is the same as the sign of A.*

Proof. Consider the objective function J, one has

$$\begin{aligned}&J(T,S,q)-J(-T,S,q)\\=&(1-S)^2+q+\frac{T^2(S^2+q)}{1-T^2}-\frac{2AT(S-ST^2-S^2-q)}{(1-T^2)(1-AT)}\\&-\left[(1-S)^2+q+\frac{T^2(S^2+q)}{1-T^2}+\frac{2AT(S-ST^2-S^2-q)}{(1-T^2)(1+AT)}\right]\\=&\frac{-2AT(S-ST^2-S^2-q)}{1-T^2}\left(\frac{1}{1-AT}+\frac{1}{1+AT}\right).\end{aligned}$$

When the sign of T is the same as the sign of A, i.e., $AT>0$, the above equation is non-negative, which implies $J(T,S,q)\ge J(-T,S,q)$. □

Lemma A.5. *The attack tuple (T^*,S^*,q^*) that maximizes the performance degradation ratio for the estimation error covariance satisfies $-\frac{1}{2}-\frac{1}{2}\log(S^{*2}+q^*)+\frac{S^{*2}+q^*}{2(1-T^{*2})}=\epsilon$, where $S^*<0$.*

Proof. First, we assume that there exists an attack tuple (T_e,S_e,q_e) such that $J(T_e,S_e,q_e)>J(T^*,S^*,q^*)$, where

$$-\frac{1}{2}-\frac{1}{2}\log(S_e^2+q_e)+\frac{S_e^2+q_e}{2(1-T_e^2)}<\epsilon. \tag{34}$$

Let S_e^* denote the corresponding smallest solution to the equation $-\frac{1}{2}-\frac{1}{2}\log(S_e^{*2}+q_e)+\frac{S_e^{*2}+q_e}{2(1-T_e^2)}=\epsilon$. Considering the derivative of J with respect to S and the property of the constraint, one can verify that $J(T_e,S_e^*,q_e)>J(T_e,S_e,q_e)$. Since among all the attack tuples satisfying the constraint equality, (T^*,S^*,q^*) is the optimal one that achieves the maximum value of J, we have $J(T^*,S^*,q^*)\ge J(T_e,S_e^*,q_e)$. Hence, $J(T^*,S^*,q^*)>J(T_e,S_e,q_e)$ is a contradiction to the early assumption. The proof is completed. □

For the simplicity of analysis, we only consider $T>0$ and $A>0$. Hence, T is non-negative in the above equation. The case when $T<0$ and $A<0$ is essentially the same.

Reconsider the constraint function of (30). Define $\mathscr{S}\triangleq S^2+q$ and

$$\mathscr{C}\triangleq -\frac{1}{2}-\frac{1}{2}\log(\mathscr{S})+\frac{\mathscr{S}}{2(1-T^2)}-\epsilon. \tag{35}$$

It is easy to obtain that $\mathscr{C}$ takes the minimum value at $\mathscr{S}=1-T^2$. Since $\mathscr{C}$ must satisfy $\mathscr{C}\le 0$, $\mathscr{S}\ge e^{-2\epsilon}$ should hold. Hence, the range of S is $-S_{oq\max}<S\le-\sqrt{e^{-2\epsilon}-q}$, where $-S_{oq\max}$ is the smaller solution to the equation $S^2+q=1+\log(S^2+q)+2\epsilon$, which implies the critical solution when $T=0$. One can prove Lemma 4.1 by the above lemmas.

Appendix B. Proof of Lemma 4.2

Compute the derivative of J_{q2}:

$$\begin{aligned}&J'_{q2}(S)\\=&(-2)\frac{SA^2f_q^2(S)+S^2+q-A(S^2+q)f_q(S)-S}{(S^2+q)(1-Af_q(S))^2}\\&-2\frac{\left[S(S^2+q)-(S^2+q)(2\epsilon+1+\log(S^2+q))\right]Af'_q(S)}{(S^2+q)(1-Af_q(S))^2},\end{aligned} \tag{36}$$

where $f'_q(S)=-\frac{\frac{S(2\epsilon+1+\log(S^2+q))-S}{(2\epsilon+1+\log(S^2+q))^2}}{\sqrt{1-\frac{S^2+q}{2\epsilon+1+\log(S^2+q)}}}$.

First we consider the left boundary. Since there is no derivative of J_{q2} at $S=-S_{oq\max}$, we consider the local property near $S=-S_{oq\max}$. Let us take $S=S_\delta$, where $\frac{S_\delta^2+q}{2\epsilon+1+\log(S_\delta^2+q)}=1-\delta$ $(0<\delta<1)$. When $\delta\to 0$, we have

$$f_q(S_\delta)=\sqrt{1-\frac{S_\delta^2+q}{2\epsilon+1+\log(S_\delta^2+q)}}=\sqrt{\delta}.$$

Hence, we rewrite the numerator of (36) as follows:

$$\begin{aligned}&\lim_{\delta\to0}S_\delta A^2f_q^2(S_\delta)+S_\delta^2+q-A(S_\delta^2+q)f_q(S_\delta)-S_\delta\\&+\left[S_\delta(S_\delta^2+q)-(S^2+q)(2\epsilon+1+\log(S_\delta^2+q))\right]Af'_q(S_\delta)\\=&\lim_{\delta\to0}S_\delta A^2\delta+S_\delta^2+q-A(S_\delta^2+q)\sqrt{\delta}-S_\delta\\&+\left[S_\delta(S_\delta^2+q)-(S_\delta^2+q)(2\epsilon+1+\log(S_\delta^2+q))\right]Af'_q(S_\delta),\\=&\lim_{\delta\to0}S_\delta^2+q-S_\delta+(S_\delta^2+q)\left(S_\delta-\frac{S_\delta^2+q}{1-\delta}\right)Af'_q(S_\delta),\end{aligned} \tag{37}$$

where $f'_q(S_\delta)=-\frac{\frac{S_\delta(2\epsilon+1+\log(S_\delta^2+q))-S_\delta}{(2\epsilon+1+\log(S_\delta^2+q))^2}}{\sqrt{1-\frac{S_\delta^2+q}{2\epsilon+1+\log(S_\delta^2+q)}}}=-\frac{S_\delta(\frac{S_\delta^2+q}{1-\delta}-1)}{(\frac{S_\delta^2+q}{1-\delta})^2\sqrt{\delta}}$.

Hence, as δ approaches to 0, (37) is given by

$$\lim_{\delta\to0}S_\delta^2+q-S_\delta-(S_\delta^2+q)\left(S_\delta-\frac{S_\delta^2+q}{1-\delta}\right)A\frac{S_\delta(\frac{S_\delta^2+q}{1-\delta}-1)}{(\frac{S_\delta^2+q}{1-\delta})^2\sqrt{\delta}}. \tag{38}$$

Since $\lim_{\delta\to0}\left[-(S_\delta^2+q)\left(S_\delta-\frac{S_\delta^2+q}{1-\delta}\right)A\frac{S_\delta}{(\frac{S_\delta^2+q}{1-\delta})^2\sqrt{\delta}}\right]=-\infty$, the sign of (38) is determined by the sign of $S_\delta^2+q-1+\delta$. Hence, we have

$$\lim_{\delta\to0}S_\delta^2+q-1+\delta=\lim_{S_\delta\to-S_{oq\max}}S_\delta^2+q-1+\delta>0.$$

Hence, when $S_\delta\to-S_{oq\max}^+$, the derivative of J_{q2} is positive.

When $S_\epsilon=-\sqrt{e^{-2\epsilon}-q}$, we have:

$$f_q(S_\epsilon)=\sqrt{1-\frac{e^{-2\epsilon}}{2\epsilon+1+\log(e^{-2\epsilon})}}=\sqrt{1-e^{-2\epsilon}},$$

and

$$f'_q(S_\epsilon)=-\frac{\frac{-\sqrt{e^{-2\epsilon}-q}(2\epsilon+1+\log(e^{-2\epsilon}))+\sqrt{e^{-2\epsilon}-q}}{(2\epsilon+1+\log(e^{-2\epsilon}))^2}}{\sqrt{1-\frac{e^{-2\epsilon}}{2\epsilon+1+\log(e^{-2\epsilon})}}}=0.$$

$$
\begin{aligned}
&S_\epsilon A^2 f_q^2(S_\epsilon) + S_\epsilon^2 + q - A(S_\epsilon^2 + q) f_q(S_\epsilon) - S_\epsilon \\
&\quad + \left[S_\epsilon(S_\epsilon^2 + q) - (S_\epsilon^2 + q)(2\epsilon + 1 + \log(S_\epsilon^2 + q))\right] A f_q'(S_\epsilon) \\
=&S_\epsilon A^2(1 - e^{-2\epsilon}) + e^{-2\epsilon} - A e^{-2\epsilon}\sqrt{1 - e^{-2\epsilon}} - S_\epsilon \\
=&S_\epsilon[A^2(1 - e^{-2\epsilon}) - 1] + e^{-2\epsilon}(1 - A\sqrt{1 - e^{-2\epsilon}}) \overset{(b)}{>} 0,
\end{aligned}
$$

where inequality (b) holds since $A^2 < 1$, $1 - e^{-2\epsilon} \leq 1$ and $S_\epsilon < 0$. Hence, the derivative of J_{q2} at $S = -\sqrt{e^{-2\epsilon} - q}$ is negative.

Since the function J_1 is continuous, there must be at least one maximum point where its first derivative is zero. Hence, $\eta = 1 + \frac{J_{q\,\mathrm{opt}} A^2 K^2 \sigma_z^2}{(1 - A^2)P}$, where $J_{q\,\mathrm{opt}} = J_{q2}(S_q)$. □

Appendix C. Proof of Lemma 4.3

By analyzing the derivative of J with respect to S, combining (23), (24), and Lemma 4.2, we know that when S takes its minimum value, J obtains the maximum. Hence, $q = 0$ performs better than $q > 0$. In other words, the solution to the optimization problem (15) requires $q = 0$. □

References

Bai, C.-Z., & Gupta, V. (2014). On Kalman filtering in the presence of a compromised sensor: Fundamental performance bounds. In *Proceedings of American control conference* (pp. 3029–3034). IEEE.

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

Bai, C.-Z., Pasqualetti, F., & Gupta, V. (2015). Security in stochastic control systems: Fundamental limitations and performance bounds. In *Proceedings of American control conference* (pp. 195–200). IEEE.

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

Chen, Y., Kar, S., & Moura, J. M. (2017). Optimal attack strategies subject to detection constraints against cyber-physical systems. *IEEE Transactions on Control of Network Systems*, *5*(3), 1157–1168.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Guo, Z., Shi, D., Johansson, K. H., & Shi, L. (2016). Optimal linear cyber-attack on remote state estimation. *IEEE Transactions on Control of Network Systems*, *4*(1), 4–13.

Guo, Z., Shi, D., Johansson, K. H., & Shi, L. (2017). Consequence analysis of innovation-based integrity attacks with side information on remote state estimation. *IFAC-PapersOnLine*, *50*(1), 8399–8404.

Guo, Z., Shi, D., Johansson, K. H., & Shi, L. (2018). Worst-case stealthy innovation-based linear attack on remote state estimation. *Automatica*, *89*, 117–124.

Guo, Z., Shi, D., Johansson, K. H., & Shi, L. (2019). Worst-case innovation-based integrity attacks with side information on remote state estimation. *IEEE Transactions on Control of Network Systems*, [ISSN: 2325-5870] *6*(1), 48–59.

Hubbard, J. H., & Hubbard, B. B. (2015). *Vector calculus, linear algebra, and differential forms: a unified approach*. Matrix Editions.

Jones, S. (2019). Venezuela blackout: what caused it and what happens next? *The Guardian*, *13*.

Karnouskos, S. (2011). Stuxnet worm impact on industrial cyber-physical system security. In *Proceedings of the 37th annual conference of the IEEE industrial electronics society* (pp. 4490–4494). IEEE.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Kung, E., Dey, S., & Shi, L. (2016). The performance and limitations of *epsilon*-stealthy attacks on higher order systems. *IEEE Transactions on Automatic Control*, *62*(2), 941–947.

Lewis, F. L. (2004). Wireless sensor networks. In *Smart environments: technologies, protocols, and applications* (pp. 11–46). Wiley Online Library.

Li, Y., Shi, L., & Chen, T. (2017). Detection against linear deception attacks on multi-sensor remote state estimation. *IEEE Transactions on Control of Network Systems*, *5*(3), 846–856.

Li, Y.-G., & Yang, G.-H. (2019). Optimal stealthy false data injection attacks in cyber-physical systems. *Information Sciences*, *481*, 474–490.

Liu, H., Ni, Y., Xie, L., & Johansson, K. H. (2020). An optimal linear attack strategy on remote state estimation. arXiv:2006.04657.

Mishra, S., Shoukry, Y., Karamchandani, N., Diggavi, S. N., & Tabuada, P. (2016). Secure state estimation against sensor attacks in the presence of noise. *IEEE Transactions on Control of Network Systems*, *4*(1), 49–59.

Mo, Y., Chabukswar, R., & Sinopoli, B. (2013). Detecting integrity attacks on scada systems. *IEEE Transactions on Control Systems Technology*, *22*(4), 1396–1407.

Mo, Y., & Sinopoli, B. (2009). Secure control against replay attacks. In *Proceedings of the 47th annual allerton conference on communication, control, and computing* (pp. 911–918). IEEE.

Mo, Y., & Sinopoli, B. (2015). On the performance degradation of cyber-physical systems under stealthy integrity attacks. *IEEE Transactions on Automatic Control*, *61*(9), 2618–2624.

Mo, Y., Weerakkody, S., & Sinopoli, B. (2015). Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs. *IEEE Control Systems Magazine*, *35*(1), 93–109.

Qin, J., Li, M., Shi, L., & Yu, X. (2018). Optimal denial-of-service attack scheduling with energy constraint over packet-dropping networks. *IEEE Transactions on Automatic Control*, *63*(6), 1648–1663.

Ribeiro, A., Giannakis, G. B., & Roumeliotis, S. I. (2006). Soi-KF: Distributed Kalman filtering with low-cost communications using the sign of innovations. *IEEE Transactions on Signal Processing*, *54*(12), 4782–4795.

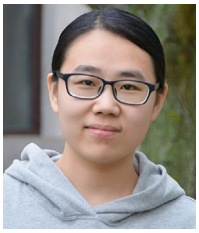
Shoukry, Y., Nuzzo, P., Puggelli, A., Sangiovanni-Vincentelli, A. L., Seshia, S. A., & Tabuada, P. (2017). Secure state estimation for cyber-physical systems under sensor attacks: A satisfiability modulo theory approach. *IEEE Transactions on Automatic Control*, *62*(10), 4917–4932.

Slay, J., & Miller, M. (2007). Lessons learned from the maroochy water breach. In *International conference on critical infrastructure protection* (pp. 73–82). Springer.

Teixeira, A., Shames, I., Sandberg, H., & Johansson, K. H. (2012). Revealing stealthy attacks in control systems. In *Proceedings of the 50th annual allerton conference on communication, control, and computing* (pp. 1806–1813). IEEE.

Whitehead, D. E., Owens, K., Gammel, D., & Smith, J. (2017). Ukraine cyber-induced power outage: Analysis and practical mitigation strategies. In *Proceedings of the 70th annual conference for protective relay engineers* (pp. 1–8). IEEE.

Zhang, H., Cheng, P., Shi, L., & Chen, J. (2015). Optimal denial-of-service attack scheduling with energy constraint. *IEEE Transactions on Automatic Control*, *60*(11), 3023–3028.



Hanxiao Liu received the B. Eng. degree from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2017. She is currently a Ph. D. student of the joint NTU-KTH programme at the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore and the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden. Her research interests include security in cyber–physical system and wireless sensor networks.

Yuqing Ni is an associate professor at the School of Internet of Things Engineering, Jiangnan University, Wuxi, China. She received the B.Eng. degree from the College of Control Science and Engineering, Zhejiang University, Hangzhou, China, and the Ph.D. degree in electronic and computer engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2016, and 2020, respectively. From April 2019 to June 2019, she was a visiting student in the Department of Automatic, KTH Royal Institute of Technology, Stockholm, Sweden. Prior to her current position, she was a senior engineer at Huawei from 2020 to 2021. Her research interests include security and privacy in cyber–physical system, networked state estimation, and wireless sensor networks.

Lihua Xie received the Ph.D. degree in electrical engineering from the University of Newcastle, Australia, in 1992. Since 1992, he has been with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he is currently a professor and Director, Delta-NTU Corporate Laboratory for Cyber–Physical Systems and Centre for Advanced Robotics Technology Innovation. He served as the Head of Division of Control and Instrumentation from July 2011 to June 2014. Dr Xie's research interests include robust control and estimation, networked control systems, multi-agent networks, localization and unmanned systems. He is an Editor-in-Chief for Unmanned Systems and has served as an editor of IET Book Series in Control and an Associate Editor of a number of journals including IEEE Transactions on Automatic Control, Automatica, IEEE Transactions on Control Systems Technology, IEEE Transactions on Control of Network Systems, and IEEE Transactions on Circuits and Systems-II. He was an elected member of Board of Governors, IEEE Control System Society (Jan 2016–Dec 2018). Dr Xie is Fellow of IEEE, IFAC and Academy of Engineering Singapore.

Karl Henrik Johansson is Professor with the School of Electrical Engineering and Computer Science at KTH Royal Institute of Technology in Sweden and Director of Digital Futures. He received M.Sc. and Ph.D. degrees from Lund University. He has held visiting positions at UC Berkeley, Caltech, NTU, HKUST Institute of Advanced Studies, and NTNU. His research interests are in networked control systems and cyber–physical systems with applications in transportation, energy, and automation networks. He is a member of the Swedish Research Council's Scientific Council for Natural Sciences and Engineering Sciences. He has served on the IEEE Control Systems Society Board of Governors, the IFAC Executive Board, and is currently Vice-President of the European Control Association. He has received several best paper awards and other distinctions from IEEE, IFAC, and ACM. He has been awarded Distinguished Professor with the Swedish Research Council and Wallenberg Scholar with the Knut and Alice Wallenberg Foundation. He has received the Future Research Leader Award from the Swedish Foundation for Strategic Research and the triennial Young Author Prize from IFAC. He is Fellow of the IEEE and the Royal Swedish Academy of Engineering Sciences, and he is IEEE Control Systems Society Distinguished Lecturer.