Data-Injection Attacks in Stochastic Control Systems: Detectability and Performance Tradeoffs ⋆

Cheng-Zong Bai [a], Fabio Pasqualetti [b], Vijay Gupta [a]

[a] *Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN USA*

[b] *Department of Mechanical Engineering, University of California, Riverside, CA USA*

Abstract

Consider a stochastic process being controlled across a communication channel. The control signal that is transmitted across the control channel can be replaced by a malicious attacker. The controller is allowed to implement any arbitrary detection algorithm to detect if an attacker is present. This work characterizes some fundamental limitations of when such an attack can be detected, and quantifies the performance degradation that an attacker that seeks to be undetected or stealthy can introduce.

Key words: Cyberphysical system security, networked control systems, stochastic systems

1 Introduction

Using communication channels to inject malicious data that degrades the performance of a cyber-physical system has now been demonstrated both theoretically and practically Farwell & Rohozinski (2011), Kuvshinkova (2003), Mo et al. (2014), Pasqualetti et al. (2013), Richards (2008), Slay & Miller (2007). Intuitively, there is a tradeoff between the performance degradation an attacker can induce and how easy it is to detect the attack Teixeira et al. (2012). Quantifying this tradeoff is of great interest to operate and design secure cyber-physical systems (CPS).

As explained in more detail later, for noiseless systems, zero dynamics provide a fundamental notion of stealthiness of an attacker, which characterizes the ability of an attacker to stay undetected even if the controller can perform arbitrary tests on the data it receives. However, similar notions for stochastic systems have been lacking. In this work, we consider stochastic cyber-physical systems, propose a graded stealthiness notion, and characterize the performance degradation that an attacker with a given level of stealthiness can induce. The proposed notion is fundamental in the sense that we do not constraint the detection test that the controller can employ to detect the presence of an attack.

Related work Security of cyber-physical systems is a growing research area. Classic works in this area focus on the detection of sensor and actuator failures in control systems Patton et al. (1989), whereas more recent approaches consider the possibility of intentional attacks at different system layers; e.g., see Pasqualetti et al. (2015). Both simple attacks, such as jamming of communication channels Foroush & Martínez (2013), and more sophisticated attacks, such as replay and data injection attacks, have been considered Mo & Sinopoli (2010), Smith (2011).

One way to organize the literature in this area is based on the properties of the considered cyber-physical systems. While initial studies focused on static systems Dan & Sandberg (2010), Giani et al. (2011), Liu et al. (2009), Mohsenian-Rad & Leon-Garcia (2011), Teixeira et al. (2010), later works exploited the dynamics of the system either to design attacks or to improve the performance of the detector that a controller can employ to detect if an attack is present Bhattacharya & Başar (2013), Hamza et al. (2011), Maharjan et al. (2013),

⋆ Work supported in part by awards NSF ECCS-1405330 and ONR N00014-14-1-0816. Corresponding author: V. Gupta. Tel. +1 574 631 2294. A preliminary version of this work appeared in Bai et al. (2015). With respect to Bai et al. (2015), this paper (i) considers more general systems with multiple inputs and outputs, (ii) completes and extends technical proofs, and (iii) provides further insight into the design of optimal stealthy attacks in stochastic cyber-physical systems.

Email addresses: cbai@nd.edu (Cheng-Zong Bai), fabiopas@engr.ucr.edu (Fabio Pasqualetti), vgupta2@nd.edu (Vijay Gupta).

Manshaei et al. (2011), Zhu & Martínez (2011), Zhu et al. (2013). For noiseless cyber-physical systems, the concept of stealthiness of an attack is closely related to the control-theoretic notion of zero dynamics (Basile & Marro 1991, Section 4). In particular, an attack is undetectable in noiseless systems if and only if it excites only the zero dynamics of an appropriately defined input-output system describing the system dynamics, the measurements available to the security monitor, and the variables compromised by the attacker Fawzi et al. (2014), Pasqualetti et al. (2013). For cyber-physical systems driven by noise, instead, the presence of process and measurements noise offers the attacker an additional possibility to tamper with sensor measurements and control inputs within acceptable uncertainty levels, thereby making the detection task more difficult.

Detectability of attacks in stochastic systems remains an open problem. Most works in this area consider detectability of attacks with respect to specific detection schemes employed by the controller, such as the classic bad data detection algorithm Cui et al. (2012), Mo & Sinopoli (2010). The trade-off between stealthiness and performance degradation induced by an attacker has also been characterized only for specific systems and detection mechanisms Kosut et al. (2011), Kwon et al. (2013), Liu et al. (2011), Mo et al. (2014), and a thorough analysis of resilience of stochastic control systems to arbitrary attacks is still missing. While convenient for analysis, the restriction to a specific class of detectors prevents the characterization of fundamental detection limitations. In our previous work Bai & Gupta (2014), we proposed the notion of ϵ-marginal stealthiness to quantify the stealthiness level in an estimation problem with respect to the class of ergodic detectors. In this work, we remove the assumption of ergodicity and introduce a notion of stealthiness for stochastic control systems that is independent of the attack detection algorithm, and thus provides a fundamental measure of the stealthiness of attacks in stochastic control systems. Further, we also characterize the performance degradation that such a stealthy attack can induce.

We limit our analysis to linear, time-invariant plants with a controller based on the output of an asymptotic Kalman filter, and to injection attacks against the actuation channel only. Our choice of using controllers based on Kalman filters is not restrictive. In fact, while this is typically the case in practice, our results and analysis are valid for arbitrary control schemes. Our choice of focusing on attacks against the actuation channel only, instead, is motivated by two main reasons. First, actuation and measurements channels are equally likely to be compromised, especially in networked control systems where communication between sensors, actuators, plant, and controller takes place over wireless channels. Second, this case has received considerably less attention in the literature – perhaps due to its enhanced difficulty – where most works focus on attacks against the measurement channel only; e.g., see Fawzi et al. (2014), Teixeira et al. (2010). We remark also that our framework can be extended to the case of attacks against the measurement channel, as we show in Bai & Gupta (2014) for scalar systems and a different notion of stealthiness.

Finally, we remark that since the submission of this work, some recent literature has appeared that builds on it and uses a notion of attack detectability that is similar to what we propose in Bai & Gupta (2014), Bai et al. (2015) and in this paper. For instance, Kung et al. (2016) extends the notion of ϵ-stealthiness of Bai et al. (2015) to higher order systems, and shows how the performance of the attacker may differ in the scalar and vector cases (in this paper we further extend the setup in Kung et al. (2016) by leveraging the notion of right-invertibility of a system to consider input and output matrices of arbitrary dimensions). In Zhang & Venkitasubramaniam (2016), the authors extend the setup in Bai et al. (2015) to vector and not necessarily stationary systems, but consider a finite horizon problem. In Guo et al. (2016), the degradation of remote state estimation is studied, for the case of an attacker that compromises the system measurements based on a linear strategy. Two other relevant recent works are Weerakkody et al. (2016) that uses the notion of Kullback-Liebler divergence as a causal measure of information flow to quantify the effect of attacks on the system output, while Chen et al. (2016) characterizes optimal attack strategies with respect to a linear quadratic cost that combines attackers control and undetectability goals.

Contributions The main contributions of this paper are threefold. First, we propose a notion of ϵ-stealthiness to quantify detectability of attacks in stochastic cyber-physical systems. Our metric is motivated by the Chernoff-Stein Lemma in detection and information theories and is universal because it is independent of any specific detection mechanism employed by the controller. Second, we provide an information theoretic bound for the degradation of the minimum-mean-square estimation error caused by an ϵ-stealthy attack as a function of the system parameters, noise statistics, and information available to the attacker. Third, we characterize optimal stealthy attacks, which achieve the maximal degradation of the estimation error covariance for a stealthy attack. For right-invertible systems (Basile & Marro 1991, Section 4.3.2), we provide a closed-form expression of optimal ϵ-stealthy attacks. The case of single-input single-output systems considered in our conference paper Bai et al. (2015) is a special case of this analysis. For systems that are not right-invertible, we propose a sub-optimal ϵ-stealthy attack with an analytical expression for the induced degradation of the system performance. We include a numerical study showing the effectiveness of our bounds. Our results provide a quantitative analysis of the trade-off between performance degradation that an attacker can induce versus a fundamental limit of the detectability of the attack.

Paper organization Section 2 contains the mathematical formulation of the problems considered in this paper. In Section 3, we propose a metric to quantify the stealthiness level of an attacker, and we characterize how this metric relates to the information theoretic notion of Kullback-Leibler Divergence. Section 4 contains the main results of this paper, including a characterization of the largest performance degradation caused by an ϵ-stealthy attack, a closed-form expression of optimal ϵ-stealthy attacks for right invertible systems, and a suboptimal class of attacks for not right-invertible systems. Section 5 presents illustrative examples and numerical results. Finally, Section 6 concludes the paper.

2 Problem Formulation

Notation: The sequence $\{x_n\}_{n=i}^{j}$ is denoted by x_i^j (when clear from the context, the notation x_i^j may also denote the corresponding vector obtained by stacking the appropriate entries in the sequence). This notation allows us to denote the probability density function of a stochastic sequence x_i^j $f_{x_i^j}$, and to define its differential entropy $h(x_i^j)$ as (Cover & Thomas 2006, Section 8.1)

$$h(x_i^j) \triangleq \int_{-\infty}^{\infty} -f_{\tilde{x}_i^j}(t_i^j) \log f_{\tilde{x}_i^j}(t_i^j) dt_i^j.$$

Let x_1^k and y_1^k be two random sequences with probability density functions (pdf) $f_{x_1^k}$ and $f_{y_1^k}$, respectively. The Kullback-Leibler Divergence (KLD) (Cover & Thomas 2006, Section 8.5) between x_1^k and y_1^k is defined as

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

The KLD is a non-negative quantity that gauges the dissimilarity between two probability density functions with $D\big(x_1^k \big\| y_1^k\big) = 0$ if $f_{x_1^k} = f_{y_1^k}$. Also, the KLD is generally not symmetric, that is, $D\big(x_1^k \big\| y_1^k\big) \neq D\big(y_1^k \big\| x_1^k\big)$. A Gaussian random vector x with mean μ_x and covariance matrix Σ_x is denoted by $x \sim \mathcal{N}(\mu_x, \Sigma_x)$. We let I and O be the identity and zero matrices, respectively, with their dimensions clear from the context. We also let $\mathbb{S}_+^n$ and $\mathbb{S}_{++}^n$ denote the sets of $n \times n$ positive semidefinite and positive definite matrices, respectively. For a square matrix M, $\mathrm{tr}(M)$ and $\det(M)$ denote the trace and the determinant of M, respectively.

We consider the setup shown in Figure 1 with the following assumptions:

Process: The process is described by the following linear time-invariant (LTI) state-space representation:

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

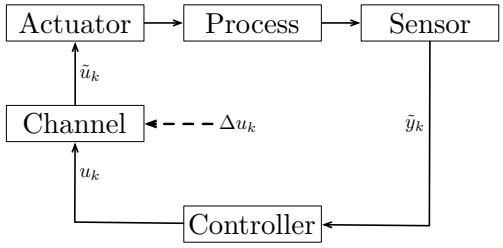

Fig. 1. Problem setup considered in the paper.

where $x_k \in \mathbb{R}^{N_x}$ is the process state, $u_k \in \mathbb{R}^{N_u}$ is the control input, $y_k \in \mathbb{R}^{N_y}$ is the output measured by the sensor, and the sequences w_1^∞ and v_1^∞ represent process and measurement noises, respectively.

Assumption 1 *The noise random processes are independent and identically distributed (i.i.d.) sequences of Gaussian random vectors with $w_k \sim \mathcal{N}(0, \Sigma_w)$, $v_k \sim \mathcal{N}(0, \Sigma_v)$, $\Sigma_w \in \mathbb{S}_{++}^{N_x}$, and $\Sigma_v \in \mathbb{S}_{++}^{N_y}$.*

Assumption 2 *The state-space realization (A, B, C) has no invariant zeros (Basile & Marro 1991, Section 4.4). In particular, this assumption implies that the system (A, B, C) is both controllable and observable.*

Assumption 3 *The controller uses a Kalman filter to estimate and monitor the process state. Note that the control input itself may be calculated using an arbitrary control law. The Kalman filter, which calculates the Minimum-Mean-Squared-Error (MMSE) estimate $\hat{x}_k$ of x_k from the measurements y_1^{k-1}, is described as*

$$\hat{x}_{k+1} = A\hat{x}_k + K_k(y_k - C\hat{x}_k) + Bu_k, \tag{3}$$

where the Kalman gain K_k and the error covariance matrix $P_{k+1} \triangleq \mathbb{E}\big[(\hat{x}_{k+1} - x_{k+1})(\hat{x}_{k+1} - x_{k+1})^T\big]$ are calculated through the recursions

$$\begin{aligned} K_k &= AP_kC^T(CP_KC^T + \Sigma_v)^{-1}, \text{ and} \\ P_{k+1} &= AP_kA^T - AP_kC^T(CP_kC^T + \Sigma_v)^{-1}CP_kA^T + \Sigma_w, \end{aligned}$$

with initial conditions $\hat{x}_1 = \mathbb{E}[x_1] = 0$ and $P_1 = \mathbb{E}[x_1 x_1^T]$.

Assumption 4 *Given Assumption 2, $\lim_{k\to\infty} P_k = P$, where P is the unique solution of a discrete-time algebraic Riccati equation. For ease of presentation, we assume that $P_1 = P$, although the results can be generalized to the general case at the expense of more involved notation. Accordingly, we drop the time index and let $K_k = K$ and $P_k = P$ at every time step k. Notice that this assumption also implies that the innovation sequence z_1^∞ calculated as $z_k \triangleq y_k - C\hat{x}_k$ is an i.i.d. Gaussian process with $z_k \sim \mathcal{N}(0, \Sigma_z)$, where $\Sigma_z = CPC^T + \Sigma_v \in \mathbb{S}_{++}^{N_y}$.*

Let $G(\mathcal{Z})$ denote the $N_y \times N_u$ matrix transfer function of the system (A, B, C). We say that the system (A, B, C) is

right invertible if there exists an $N_u \times N_y$ matrix transfer function $G_{RI}(\mathcal{Z})$ such that $G(\mathcal{Z})G_{RI}(\mathcal{Z}) = I_{N_y}$.

Attack model: An attacker can replace the input sequence u_1^∞ with an arbitrary sequence $\tilde{u}_1^\infty$. Thus, in the presence of an attack, the system dynamics are given by

$$\begin{aligned} \tilde{x}_{k+1} &= A\tilde{x}_k + B\tilde{u}_k + w_k, \\ \tilde{y}_k &= C\tilde{x}_k + v_k. \end{aligned} \tag{4}$$

Note that the sequence $\tilde{y}_1^\infty$ generated by the sensor in the presence of an attack $\tilde{u}_1^\infty$ is different from the nominal measurement sequence y_1^∞. We assume that the attacker knows the system parameters, including the matrices A, B, C, Σ_w, and Σ_v. The attack input $\tilde{u}_1^\infty$ is constructed based on the system parameters and the *information pattern* $\mathcal{I}_k$ of the attacker. We make the following assumptions on the attacker's information pattern:

Assumption 5 *The attacker knows the control input u_k; thus $u_k \in \mathcal{I}_k$ at all times k. Additionally, the attacker does not know the noise vectors for any time.*

Assumption 6 *The attacker has perfect memory; thus, $\mathcal{I}_k \subseteq \mathcal{I}_{k+1}$ at all times k.*

Assumption 7 *The attacker has causal information; in particular, $\mathcal{I}_k$ is independent of w_k^∞ and v_{k+1}^∞ for all k.*

Example* 1** ***(Attack scenarios) *Attack scenarios satisfying Assumptions 5-7 include the cases when:*

(i) the attacker knows the control input exactly, that is, $\mathcal{I}_k = \{u_1^k\}$.
(ii) the attacker knows the control input and the state, that is, $\mathcal{I}_k = \{u_1^k, x_1^k\}$.
(iii) the attacker knows the control input and delayed measurements from the sensor, that is, $\mathcal{I}_k = \{u_1^k, \tilde{y}_1^{k-d}\}$ for some $d \geq 1$.

Stealthiness of an attacker: The attacker is constrained in the input $\tilde{u}_1^\infty$ it replaces since it seeks to be stealthy or undetected by the controller. If the controller is aware that an attacker has replaced the correct control sequence u_1^∞ by a different sequence $\tilde{u}_1^\infty$, it can presumably switch to a safer mode of operation. Notions of stealthiness have been proposed in the literature before. As an example, for noiseless systems, Pasqualetti et al. (2013) showed that stealthiness of an attacker is equivalent to the existence of zero dynamics for the system driven by the attack. Similar to Pasqualetti et al. (2013), we seek to define the notion of stealthiness without placing any restrictions on the attacker or the controller behavior. However, we need to define a similar notion for stochastic systems when zero dynamics may not exist. To this end, we pose the problem of detecting an attacker by the controller as a (sequential) hypothesis testing problem. Specifically, the controller relies on the received measurements to decide the following binary hypothesis testing problem:

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

For a given detector employed at the controller to select one of the two hypotheses, denote the probability of false alarm (i.e., the probability of deciding H_1 when H_0 is true) at time k by p_k^F, and the probability of correct detection (i.e., the probability of deciding H_1 when H_1 is true) at time $k+1$ by p_k^D.

One may envisage that stealthiness of an attacker implies $p_k^D = 0$. However, as is standard in detection theory, we need to consider both p_k^F and p_k^D simultaneously. For instance, a detector that always declares H_1 to be true will achieve $p_k^D = 1$. However, it will not be a good detector because $p_k^F = 1$. Intuitively, an attack is harder to detect if the performance of *any* detector is independent of the received measurements. In other words, we define an attacker to be stealthy if there exists no detector that can perform better (in the sense of simultaneously achieving higher p_k^D and lower p_k^F) than a detector that makes a decision by ignoring all the measurements and making a random guess to decide between the hypotheses. We formalize this intuition in the following definition.

Definition* 1** ***(Stealthy attacks) *Consider the problem formulation stated in Section 2. An attack $\tilde{u}_1^\infty$ is*

(i) strictly stealthy, if there exists no detector such that $p_k^F < p_k^D$ for any $k > 0$.
(ii) ϵ-stealthy, if, given $\epsilon > 0$ and for any $0 < \delta < 1$, for any detector for which $0 < 1 - p_k^D \leq \delta$ for all times k, it holds that $\limsup_{k\to\infty} -\frac{1}{k} \log p_k^F \leq \epsilon$.

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

Performance metric: The requirement to stay stealthy clearly curtails the performance degradation that an attacker can cause. The central problem that we consider is to characterize the worst performance degradation that an attacker can achieve for a specified level of stealthiness. In the presence of an attack (and if the controller is unaware of the attack), it uses the corrupted measurements $\tilde{y}_1^\infty$ in the Kalman filter. Let $\hat{\tilde{x}}_1^\infty$ be the estimate of the Kalman filter (3) in the presence of the attack $\tilde{u}_1^\infty$, which is obtained from the recursion

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

where the innovation is $\tilde{z}_k \triangleq \tilde{y}_k - C\hat{\tilde{x}}_k$. Note that the estimate $\hat{\tilde{x}}_{k+1}$ is a sub-optimal MMSE estimate of the state x_k since it is obtained by assuming the nominal control input u_k, whereas the system is driven by the attack input $\tilde{u}_k$. Also, note that the random sequence $\tilde{z}_1^\infty$ need neither be zero mean, nor white or Gaussian.

Since the Kalman filter estimate depends on the measurement sequence received, as a performance metric, we consider the covariance of the error in the predicted measurement $\hat{\tilde{y}}_k$ as compared to true value y_k. Further, to normalize the relative impact of the degradation induced by the attacker among different components of this error vector, we weight each component of the error vector by an amount corresponding to how accurate the estimate of this component was without attacks. Thus, we consider the performance index $\mathbb{E}\left[\left(\hat{\tilde{y}}_k - y_k\right)^T \Sigma_z^{-1} \left(\hat{\tilde{y}}_k - y_k\right)\right] = \mathrm{Tr}(\tilde{P}_k W)$, where $\tilde{P}_k$ is the error covariance matrix in the presence of an attack, $\tilde{P}_k = \mathbb{E}\big[(\hat{\tilde{x}}_k - x_k)(\hat{\tilde{x}}_k - x_k)^T\big]$, and $W = C^T \Sigma_z^{-1} C$. To obtain a metric independent of time and focus on the long term effect of the attack, we consider the limit superior of the arithmetic mean of $\{\mathrm{tr}(\tilde{P}_k W)\}_{k=1}^\infty$ and define $\tilde{P}_W \triangleq \limsup_{k\to\infty} \frac{1}{k} \sum_{n=1}^k \mathrm{tr}(\tilde{P}_n W)$. If $\{\mathrm{tr}(\tilde{P}_k W)\}_{k=1}^\infty$ is convergent, then $\lim_{k\to\infty} \mathrm{tr}(\tilde{P}_k W) = \tilde{P}_W$, which equals the Cesàro mean of $\tilde{P}_k W$.

Problems considered in the paper: We assume that the attacker is interested in staying stealthy or undetected for as long as possible while maximizing the error covariance $\tilde{P}_W$. We consider two problems:

(i) What is a suitable metric for stealthiness of an attacker in stochastic systems where Assumption 2 holds? We consider this problem in Section 3.

(ii) For a specified level of stealthiness, what is the worst performance degradation that an attacker can achieve? We consider this problem in Section 4.

3 Stealthiness in Stochastic systems

Our first result provides conditions that can be used to verify if an attack is stealthy or not.

Theorem 1 ***(KLD and stealthy attacks)*** *Consider the problem formulation in Section 2. An attack $\tilde{u}_1^\infty$ is*

(i) strictly stealthy if and only if $D\big(\tilde{y}_1^k \big\| y_1^k\big) = 0 \; \forall k > 0$.
(ii) ϵ-stealthy if the corresponding observation sequence $\tilde{y}_1^\infty$ is ergodic and satisfies

$$\lim_{k\to\infty} \frac{1}{k} D\big(\tilde{y}_1^k \big\| y_1^k\big) \le \epsilon. \tag{5}$$

(iii) ϵ-stealthy only if the corresponding observation sequence $\tilde{y}_1^\infty$ satisfies (5).

PROOF. Presented in Appendix A. □

The following result provides a characterization of $D\big(\tilde{y}_1^k \big\| y_1^k\big)$ that contains additional insight into the meaning of stealthiness of an attacker.

Proposition 2 ***(KLD and differential entropy)*** *The quantity $D\big(\tilde{y}_1^k \big\| y_1^k\big)$ can be calculated as*

$$\frac{1}{k} D\big(\tilde{y}_1^k \big\| y_1^k\big) = \frac{1}{k} \sum_{n=1}^k \Big(I\big(\tilde{z}_1^{n-1}; \tilde{z}_n\big) + D\big(\tilde{z}_n \big\| z_n\big)\Big), \tag{6}$$

where $I\big(\tilde{z}_1^{n-1}; \tilde{z}_n\big)$ denotes the mutual information between $\tilde{z}_1^{n-1}$ and $\tilde{z}_n$ (Cover & Thomas 2006, Section 8.5).

PROOF. Due to the invariance property of the Kullback-Leibler divergence Kullback (1997), we have $D\big(\tilde{y}_1^k \big\| y_1^k\big) = D\big(\tilde{z}_1^k \big\| z_1^k\big)$, for every $k > 0$. Further, note that z_1^∞ is an i.i.d. sequence of Gaussian random vectors with $z_k \sim \mathcal{N}(0, \Sigma_z)$. From (1), we obtain

$$\begin{aligned}
\frac{1}{k} D\big(\tilde{z}_1^k \big\| z_1^k\big) &\overset{(a)}{=} -\frac{1}{k} h\big(\tilde{z}_1^k\big) - \frac{1}{k} \sum_{n=1}^k \mathbb{E}\big[\log f_{z_n}(z_n)\big] \\
&\overset{(b)}{=} \frac{1}{k} \sum_{n=1}^k \Big(- h\big(\tilde{z}_n \big| \tilde{z}_1^{n-1}\big) + h(\tilde{z}_n) \\
&\qquad - h(\tilde{z}_n) - \mathbb{E}\big[\log f_{z_n}(z_n)\big]\Big) \\
&= \frac{1}{k} \sum_{n=1}^k \Big(I\big(\tilde{z}_1^{n-1}; \tilde{z}_n\big) + D\big(\tilde{z}_n \big\| z_n\big)\Big),
\end{aligned}$$

where $I\big(\tilde{z}_1^{n-1}; \tilde{z}_n\big)$ denotes the mutual information between $\tilde{z}_1^{n-1}$ and $\tilde{z}_n$. Equality (a) holds because z_1^∞ is an independent random sequence, while (b) follows by applying the chain rule of differential entropy (Cover & Thomas 2006, Theorem 8.6.2) on the term $-\frac{1}{k} h\left(\tilde{z}_1^k\right)$ to obtain $\frac{1}{k} \sum_{n=1}^k -h\left(\tilde{z}_n | \tilde{z}_1^{n-1}\right)$, and adding and subtracting $h(\tilde{z}_n)$. □

Intuitively, the mutual information $I\big(\tilde{z}_1^{n-1}; \tilde{z}_n\big)$ measures how much information about $\tilde{z}_n$ can be obtained from $\tilde{z}_1^{n-1}$, that is, it characterizes the memory of the sequence $\tilde{z}_1^\infty$. Similarly, the Kullback-Leibler divergence $D(\tilde{z}_n \| z_n)$ measures the dissimilarity between the marginal distributions of $\tilde{z}_n$ and z_n. Proposition 2 thus states that the stealthiness level of an ergodic attacker can be degraded in two ways: (i) if the sequence $\tilde{z}_1^\infty$

becomes autocorrelated, and (ii) if the marginal distributions of the random variables $\tilde{z}(k)$ in the sequence $\tilde{z}_1^\infty$ deviate from $\mathcal{N}(0, \Sigma_z)$.

4 Fundamental Performance Limitations

We are interested in the maximal performance degradation $\tilde{P}_W$ that an ϵ-stealthy attacker may induce. We begin by proving a converse statement that gives an upper bound for $\tilde{P}_W$ induced by an ϵ-stealthy attacker in Section 4.1. In Section 4.2 we prove a tight achievability result that provides an attack that achieves the upper bound when the system (A, B, C) is right-invertible. In Section 4.3 we prove a looser achievability result that gives a lower bound on the performance degradation for non right-invertible systems.

We will use a series of preliminary technical results to present the main results of the paper. The following result is immediate.

Lemma 3 *Define the function $\bar{\delta} : [0, \infty) \to [1, \infty)$ as $\bar{\delta}(x) = 2x + 1 + \log \bar{\delta}(x)$. Then, for any $\gamma > 0$, $\bar{\delta}(\gamma) = \arg\max_{x \in \mathbb{R}} \; x$, subject to $\frac{1}{2}x - \gamma - \frac{1}{2} \leq \frac{1}{2}\log x$.*

The following result is proved in the appendix B.

Lemma 4 *Consider the problem setup above. We have*

$$\frac{1}{2k}\sum_{n=1}^{k} tr\big(\mathbb{E}\big[\tilde{z}_n\tilde{z}_n^T\big]\Sigma_z^{-1}\big) \leq \frac{N_y}{2} + \frac{1}{k}D\big(\tilde{z}_1^k\big\|z_1^k\big) + \frac{N_y}{2}\log\left(\frac{1}{N_y k}\sum_{n=1}^{k} tr\big(\mathbb{E}[\tilde{z}_n\tilde{z}_n^T]\Sigma_z^{-1}\big)\right). \quad (7)$$

Further, if the sequence $\tilde{z}_1^\infty$ is a sequence of independent and identically distributed (i.i.d.) Gaussian random variables, $\tilde{z}_k$, each with mean zero and covariance matrix $\mathbb{E}\big[\tilde{z}_k\tilde{z}_k^T\big] = \alpha\Sigma_z$, for some scalar α, then (7) is satisfied with equality.

Combining Lemmas 3 and 4 leads to the following result.

Lemma 5 *Consider the problem setup above. We have*

$$\frac{1}{N_y k}\sum_{n=1}^{k} tr\big(\mathbb{E}\big[\tilde{z}_n\tilde{z}_n^T\big]\Sigma_z^{-1}\big) \leq \bar{\delta}\Big(\frac{1}{N_y k}D\big(\tilde{z}_1^k\big\|z_1^k\big)\Big), \quad (8)$$

where $\bar{\delta}(.)$ is as defined in Lemma 3.

The following result relates the covariance of the innovation and the observation sequence.

Lemma 6 *Consider the problem setup above. We have*

$$CP_kC^T = \mathbb{E}\big[z_kz_k^T\big] - \Sigma_v \quad (9)$$
$$C\tilde{P}_kC^T = \mathbb{E}\big[\tilde{z}_k\tilde{z}_k^T\big] - \Sigma_v. \quad (10)$$

PROOF. By definition, $z_k = y_k - C\hat{x}_k = C(x_k - \hat{x}_k) + v_k$, and similarly $\tilde{z}_k = C(\tilde{x}_k - \hat{\tilde{x}}_k) + v_k$. Since $(x_k - \hat{x}_k)$ and $(\tilde{x}_k - \hat{\tilde{x}}_k)$ are independent of the measurement noise v_k due to Assumptions 1 and 7, the result follows. □

4.1 Converse

We now present an upper bound of the weighted MSE induced by an ϵ-stealthy attack.

Theorem 7 (*Converse*) *Consider the problem setup above. For any ϵ-stealthy attack $\tilde{u}_1^\infty$ generated by an information pattern $\mathcal{I}_1^\infty$ that satisfies Assumptions 5-7,*

$$\tilde{P}_W \leq tr(PW) + \left(\bar{\delta}\left(\frac{\epsilon}{N_y}\right) - 1\right)N_y, \quad (11)$$

where N_y is the number of outputs of the system, the function $\bar{\delta}$ is defined in Lemma 3, and $tr(PW)$ is the weighted MSE in the absence of the attacker.

PROOF. We begin by writing

$$\begin{aligned}\tilde{P}_W &= \limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\mathrm{tr}(\tilde{P}_nC^T\Sigma_z^{-1}C)\\ &= \limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\mathrm{tr}(C\tilde{P}_nC^T\Sigma_z^{-1})\\ &= \limsup_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\mathrm{tr}\Big(\big(\mathbb{E}[\tilde{z}_n\tilde{z}_n^T] - \Sigma_v\big)\Sigma_z^{-1}\Big),\end{aligned}$$

where we have used the invariance of trace operator under cyclic permutations and the relation in (10), respectively. The right hand side has two terms. The first term can be upper bounded using Lemma 5, so that we obtain

$$\tilde{P}_W \leq \limsup_{k\to\infty} N_y\bar{\delta}\left(\frac{1}{N_y k}D\big(\tilde{z}_1^k\big\|z_1^k\big)\right) - \mathrm{tr}\big(\Sigma_v\Sigma_z^{-1}\big).$$

Since the function $\bar{\delta}$ is continuous and monotonic, we can rewrite the above bound as

$$\tilde{P}_W \leq N_y\bar{\delta}\left(\limsup_{k\to\infty}\frac{1}{N_y k}D\big(\tilde{z}_1^k\big\|z_1^k\big)\right) - \mathrm{tr}\big(\Sigma_v\Sigma_z^{-1}\big).$$

Since the attack is ϵ-stealthy, we use Theorem 1 to bound the Kullback-Leibler divergence $D\big(\tilde{z}_1^k\big\|z_1^k\big)$ to

obtain $\tilde{P}_W \leq N_y \bar{\delta}\Big(\frac{\epsilon}{N_y}\Big) - \mathrm{tr}\big(\Sigma_v \Sigma_z^{-1}\big)$. Finally, substituting for Σ_v from (9) on the right hand side and using $W = C^T \Sigma_z^{-1} C$ completes the proof. □

Remark 8 ***(Stealthiness vs induced error)*** *Theorem 7 provides an upper bound for the performance degradation $\tilde{P}_W$ for ϵ-stealthy attacks. Since $\bar{\delta}\Big(\frac{\epsilon}{N_y}\Big)$ is a monotonically increasing function of ϵ, the upper bound (11) characterizes a trade-off between the induced error and the stealthiness level of an attack.*

To further understand this result, we consider two extreme cases, namely, $\epsilon = 0$, which implies strictly stealthiness, and $\epsilon \to \infty$, that is, no stealthiness level.

Corollary 9 *A strictly stealthy attacker cannot induce any performance degradation. Further, for an ϵ-stealthy attacker, the upper bound in (11) increases linearly with ϵ as $\epsilon \to \infty$.*

PROOF. A strictly stealthy attacker corresponds to $\epsilon = 0$. Using the fact that $\bar{\delta}(0) = 1$ in Theorem 7 yields that $\mathrm{tr}(\tilde{P}W) \leq \mathrm{tr}(PW)$. The second statement follows by noting that the first order derivative of the function $\bar{\delta}(x) \to 2$ from the right as x tends to infinity. □

4.2 Achievability for Right Invertible Systems

We now show that the bound presented in Theorem 7 is achievable if the system (A, B, C) is right invertible. We begin with the following preliminary result.

Lemma 10 *Let the system (A, B, C) be right invertible. Then, the system $(A - KC, B, C)$ is also right invertible.*

Let G'_{RI} be the right inverse of the system $(A - KC, B, C)$. We consider the following attack.

Attack $\mathcal{A}_1$: The attack sequence is generated in three steps. In the first step, a sequence ζ_1^∞ is generated, such that each vector ζ_k is independent and identically distributed and independent of the information pattern $\mathcal{I}_k$ of the attacker, with probability density function $\zeta_k \sim \mathcal{N}\big(0, \big(\bar{\delta}(\frac{\epsilon}{N_y}) - 1\big)\Sigma_z\big)$. In the second step, the sequence ϕ_1^∞ is generated as the output of the system G'_{RI} with ζ_1^∞ as the input sequence. Finally, the attack sequence $\tilde{u}_1^\infty$ is generated as $\tilde{u}_k = u_k + \phi_k$.

Remark 11 ***(Information pattern of attack $\mathcal{A}_1$)*** *The attack $\mathcal{A}_1$ can be generated by an attacker with any information pattern satisfying Assumptions 5–7.*

We note the following property of the attack $\mathcal{A}_1$.

Lemma 12 *Consider the attack $\mathcal{A}_1$. With this attack, the innovation sequence $\tilde{z}_1^\infty$ as calculated at the controller, is a sequence of independent and identically distributed Gaussian random vectors with mean zero and covariance matrix $\mathbb{E}[\tilde{z}_k \tilde{z}_k^T] = \bar{\delta}\Big(\frac{\epsilon}{N_y}\Big)\Sigma_z$.*

PROOF. Consider an auxiliary Kalman filter that is implemented as the recursion

$$\hat{x}^a_{k+1} = A\hat{x}^a_k + K z^a_k + B\tilde{u}_k, \tag{12}$$

with the initial condition $\hat{x}^a_1 = 0$ and the innovation $z^a_k = \tilde{y}_k - C\hat{x}^a_k$. The innovation sequence is independent and identically distributed with each $z^a_k \sim \mathcal{N}(0, \Sigma_z)$. Now, we express $\tilde{z}_k = z^a_k - C\tilde{e}_k$, where $\tilde{e}_k \triangleq \hat{\tilde{x}}_k - \hat{x}^a_k$. Further, $\tilde{e}_k$ evolves according to the recursion

$$\begin{aligned}\tilde{e}_{k+1} &= (A\hat{\tilde{x}}_k + K\tilde{z}_k + Bu_k) - (A\hat{x}^a_k + Kz^a_k + B\tilde{u}_k)\\ &= (A - KC)\tilde{e}_k - B\phi_k,\end{aligned} \tag{13}$$

with the initial condition $\tilde{e}_1 = 0$. Together, $\tilde{z}_k$ and (13) define a system of the form

$$\begin{aligned}\tilde{e}_{k+1} &= (A - KC)\tilde{e}_k + B(-\phi_k),\\ z^a_k - \tilde{z}_k &= C\tilde{e}_k.\end{aligned} \tag{14}$$

We now note that (i) the above system is $(A-KC, B, C)$, (ii) ϕ_1^∞ is the output of the right inverse system of $(A - KC, B, C)$ with input ζ_1^∞, and (iii) the system in equation (14) is linear. These three facts together imply that the output of (14), i.e., $\{z^a_k - \tilde{z}_k\}_{k=1}^\infty$ is a sequence of independent and identically distributed random variables with each random variable distributed as $\mathcal{N}\big(0, \big(\bar{\delta}(\frac{\epsilon}{N_y}) - 1\big)\Sigma_z\big)$. Now since z^a_k is independent of $\tilde{e}_1^k$, we obtain that $\tilde{z}_1^\infty$ is an independent and identically distributed sequence with each random variable $\tilde{z}_k$ as Gaussian with mean zero and covariance matrix $\mathbb{E}[\tilde{z}_k \tilde{z}_k^T] = \Big(\bar{\delta}\Big(\frac{\epsilon}{N_y}\Big) - 1\Big)\Sigma_z + \Sigma_z = \bar{\delta}\Big(\frac{\epsilon}{N_y}\Big)\Sigma_z$. □

Theorem 13 ***(Achievability for right invertible systems)*** *Suppose that the LTI system (A, B, C) is right invertible. The attack $\mathcal{A}_1$ is ϵ-stealthy and achieves*

$$\tilde{P}_W = tr(PW) + N_y\left(\bar{\delta}\left(\frac{\epsilon}{N_y}\right) - 1\right),$$

where $W = C^T \Sigma_z^{-1} C$.

PROOF. For the attack $\mathcal{A}_1$, Lemma 12 states that $\tilde{z}_1^\infty$ is a sequence of independent and identically distributed (i.i.d.) Gaussian random variables $\tilde{z}_k$ each with mean zero and covariance matrix $\mathbb{E}\big[\tilde{z}_k \tilde{z}_k^T\big] = \alpha\Sigma_z$, with $\alpha = \bar{\delta}(\frac{\epsilon}{N_y})$. Lemma 4, thus, implies that (7) holds with

equality. Further, following the proof of Theorem 7, if (7) holds with equality, then (11) also holds with equality. Thus, the attack $\mathcal{A}_1$ achieves the converse in terms of performance degradation.

Next we show that the attack is ϵ-stealthy. Once again, from Lemma 4 and the expression for the covariance matrix of $\tilde{z}_k$, we have for every $k > 0$,

$$\begin{aligned}\frac{1}{k}D\big(\tilde{z}_1^k\|z_1^k\big) &= \frac{1}{2k}\sum_{n=1}^{k}\operatorname{tr}\big(\mathbb{E}\big[\tilde{z}_n\tilde{z}_n^T\big]\Sigma_z^{-1}\big) - \frac{N_y}{2} \\ &\quad - \frac{N_y}{2}\log\left(\frac{1}{N_y k}\sum_{n=1}^{k}\operatorname{tr}\big(\mathbb{E}[\tilde{z}_n\tilde{z}_n^T]\Sigma_z^{-1}\big)\right) \\ &= \frac{1}{2k}\sum_{n=1}^{k}\operatorname{tr}\Big(\bar{\delta}(\frac{\epsilon}{N_y})\Sigma_z\Sigma_z^{-1}\Big) - \frac{N_y}{2} \\ &\quad - \frac{1}{2k}\sum_{n=1}^{k}\log\det\big(\bar{\delta}(\frac{\epsilon}{N_y})\Sigma_z\Sigma_z^{-1}\big) \\ &= \frac{N_y}{2}\bar{\delta}\Big(\frac{\epsilon}{N_y}\Big) - \frac{N_y}{2} - \frac{N_y}{2}\log\bar{\delta}\Big(\frac{\epsilon}{N_y}\Big) = \epsilon.\end{aligned}$$

Now with this attack, $\tilde{z}_1^\infty$ is an independent and identically distributed sequence and the measurement sequence $\tilde{y}_1^\infty$ is ergodic. Thus, from Theorem 1, the attack $\mathcal{A}_1$ is ϵ-stealthy. □

Remark 14 ***(Attacker information pattern)*** *Intuitively, we may expect that the more information about the state variables that an attacker has, larger the performance degradation it can induce. However, Theorem 7 and Theorem 13 imply that the only critical piece of information for the attacker to launch an optimal attack is the nominal control input u_1^∞.*

4.3 Achievability if System is not Right Invertible

If the system is not right invertible, the converse result in Theorem 7 may not be achieved. We now construct a heuristic attack $\mathcal{A}_2$ that allows us to derive a lower bound for the performance degradation $\tilde{P}_W$ induced by ϵ-stealthy attacks against such systems.

Attack $\mathcal{A}_2$: The attack sequence is generated as $\tilde{u}_k = u_k + L\tilde{e}_k - \zeta_k$, where $\tilde{e}_k = \hat{\tilde{x}}_k - \hat{x}_k^a$ as in (14), and the sequence ζ_1^∞ is generated such that each vector ζ_k is independent and identically distributed with probability density function $\zeta_k \sim \mathcal{N}\big(0, \Sigma_\zeta\big)$ and independent of the information pattern $\mathcal{I}_k$ of the attacker. The feedback matrix L and the covariance matrix Σ_ζ are determined in three steps, which are detailed next.

Step 1 (Limiting the memory of the innovation sequence $\tilde{z}_1^\infty$): Notice that, with the attack $\mathcal{A}_2$ and the notation in (12), the dynamics of $\tilde{e}_k$ and $\tilde{z}_k$ are given by

$$\begin{aligned}\tilde{e}_{k+1} &= (A - KC - BL)\tilde{e}_k + B\zeta_k \\ \tilde{z}_k &= C\tilde{e}_k + z_k^a.\end{aligned} \tag{15}$$

The feedback matrix L should be selected to eliminate the memory of the innovation sequence computed at the controller. One way to achieve this aim is to set $A-KC-BL=0$. In other words, if $A-KC-BL=0$, then $\tilde{z}_1^\infty$ is independent and identically distributed. It may not be possible to select L to achieve this aim exactly. Thus, we propose the following heuristic. Note that if $A - KC - BL = 0$, then the cost function $\lim_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\operatorname{tr}\big(\mathbb{E}[\tilde{e}_n\tilde{e}_n^T]W\big)$, is minimized, with $W = C^T\Sigma_z^{-1}C$. Since $\sum_{n=1}^{k}\operatorname{tr}\big(\mathbb{E}[\tilde{e}_n\tilde{e}_n^T]W\big) = \mathbb{E}\left[\sum_{n=1}^{k}\tilde{e}_n^TW\tilde{e}_n\right]$, selecting L to satisfy the constraint $A-KC-BL=0$ is equivalent to selecting L to solve a cheap Linear Quadratic Gaussian (LQG) problem (Hespanha 2009, Section VI). Thus, heuristically, we select the attack matrix L as the solution to this cheap LQG problem and, specifically, as

$$L = \lim_{\eta\to 0}(B^TT_\eta B + \eta I)^{-1}B^TT_\eta(A-KC), \tag{16}$$

where T_η is the solution to the discrete algebraic Riccati equation

$$\begin{aligned}&T_\eta = \\ &(A-KC)^T\Big(T_\eta - T_\eta B(B^TT_\eta B+\eta I)^{-1}B^TT_\eta\Big)(A-KC)+W.\end{aligned}$$

Step 2 (Selection of the covariance matrix Σ_ζ): Notice that the selection of the feedback matrix L in Step 1 is independent of the covariance matrix Σ_ζ. As the second step, we select the covariance matrix Σ_ζ such that $C\Sigma_{\tilde{e}}C^T$ is close to a scalar multiplication of Σ_z, say $\alpha^2\Sigma_z$. From (15), notice that $\lim_{k\to\infty}\mathbb{E}[\tilde{z}_k\tilde{z}_k^T] = C\Sigma_{\tilde{e}}C^T + \Sigma_z$, where $\Sigma_{\tilde{e}} \in \mathbb{S}_+^{N_x}$ is the positive semi-definite solution to the equation

$$\Sigma_{\tilde{e}} = (A - KC - BL)\Sigma_{\tilde{e}}(A - KC - BL)^T + B\Sigma_\zeta B^T. \tag{17}$$

We derive an expression for Σ_ζ from (17) by using the pseudoinverse matrices of B and C, i.e.,

$$\begin{aligned}\Sigma_\zeta = \alpha^2 B^\dagger\Big(&C^\dagger\Sigma_z(C^T)^\dagger + \\ &- (A - KC - BL)C^\dagger\Sigma_z(C^T)^\dagger(A - KC - BL)^T\Big)(B^T)^\dagger,\end{aligned} \tag{18}$$

where † denotes the pseudoinverse operation. It should be noted that the right-hand side of (18) may not be positive semidefinite. Many choices are possible to construct a positive semi-definite Σ_ζ. We propose that if the the right-hand side is indefinite, we set its negative eigenvalues to zero without altering its eigenvectors.

Step 3 (Enforcing the stealthiness level): The covariance matrix Σ_ζ obtained in Step 2 depends on the parameter α. We now select α so as to make the attack $\mathcal{A}_2$ ϵ-stealthy. To this aim, we first compute an explicit expression for the stealthiness level and the error induced by $\mathcal{A}_2$. For the entropy rate of $\tilde{z}_1^\infty$, since $\tilde{z}_1^\infty$ is Gaussian, we obtain

$$\lim_{k\to\infty}\frac{1}{k}h(\tilde{z}_1^k) = \lim_{k\to\infty} h(\tilde{z}_{k+1}|\tilde{z}_1^k) \tag{19}$$

$$= \lim_{k\to\infty}\frac{1}{2}\log\Big((2\pi e)^{N_y}\det\big(\mathbb{E}[(\tilde{z}_{k+1}-g_k(\tilde{z}_1^k)) (\tilde{z}_{k+1}-g_k(\tilde{z}_1^k))^T]\big)\Big) \tag{20}$$

$$= \frac{1}{2}\log\big((2\pi e)^{N_y}\det(CSC^T+\Sigma_z)\big) \tag{21}$$

where $g_k(\tilde{z}_1^k)$ is the minimum mean square estimate of $\tilde{e}_{k+1}$ from $\tilde{z}_1^k$, which can be obtained from Kalman filtering, and $S\in\mathbb{S}_+^{N_y}$ is the positive semidefinite solution to the following discrete algebraic Riccati equation

$$S = (A-KC-BL)\Big(S-SC^T(CSC^T+\Sigma_z)^{-1}CS\Big) \times (A-KC-BL)^T + B\Sigma_\zeta B^T. \tag{22}$$

Note that the equality (19) is due to (Cover & Thomas 2006, Theorem 4.2.1); (20) is a consequence of the maximum differential entropy lemma (Gamal & Kim 2011, Section 2.2); the positive semidefinite matrix S that solves (22) represents the steady-state error covariance matrix of the Kalman filter that estimates $\tilde{z}_{k+1}$ from $\tilde{z}_1^k$. Thus, the level of stealthiness for the attack $\mathcal{A}_2$ is

$$\lim_{k\to\infty}\frac{1}{k}D(\tilde{z}_1^k\|z_1^k) = \epsilon = -\frac{1}{2}\log((2\pi e)^{N_y}\det(CSC^T+\Sigma_z) + \frac{1}{2}\log\big((2\pi)^{N_y}\det(\Sigma_z)\big) + \frac{1}{2}\mathrm{tr}\big((C\Sigma_{\tilde{e}}C^T+\Sigma_z)\Sigma_z^{-1}\big)$$
$$= -\frac{1}{2}\log\det(I+SW) + \frac{1}{2}\mathrm{tr}(\Sigma_{\tilde{e}}W) + \frac{1}{2}N_y, \tag{23}$$

where $W = C^T\Sigma_z^{-1}C$. To conclude our design of the attack $\mathcal{A}_2$, we use (23) to solve for the desired value of α, and compute the error induced by $\mathcal{A}_2$ as

$$\tilde{P}_W = \lim_{k\to\infty}\frac{1}{k}\sum_{n=1}^{k}\mathrm{tr}(\mathbb{E}[\tilde{z}_n\tilde{z}_n^T]\Sigma_z^{-1}) - \mathrm{tr}(\Sigma_v\Sigma_z^{-1}) = \mathrm{tr}(PW) + \mathrm{tr}(\Sigma_{\tilde{e}}W) - N_y \tag{24}$$

where $\Sigma_{\tilde{e}}$ is the solution to the Lyapunov equation (17).

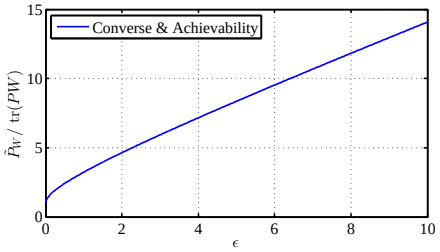

Fig. 2. The converse and achievability for the right invertible system, where the weighted MSE $\tilde{P}_W$ is the upper bound in (11) and the weight matrix $W = C^T\Sigma_z^{-1}C$.

5 Numerical Results

Example 1 Consider a right invertible system (A, B, C)

$$A = \begin{bmatrix} 2 & 0 & 0 & 0\\ 0 & -1 & 0 & 0\\ 1 & 0 & 1 & 0\\ 0 & 0 & 0 & 2\end{bmatrix}, B = \begin{bmatrix} 1 & 0\\ 1 & 0\\ 0 & 2\\ 0 & 1\end{bmatrix}, C = \begin{bmatrix} 0 & 0\\ 0 & 1\\ 2 & 0\\ 0 & 1\end{bmatrix}^T,$$

and let $\Sigma_w = 0.5I$ and $\Sigma_v = I$. Figure 2 plots the upper bound (11) of performance degradation achievable for an attacker versus the attacker's stealthiness level ϵ. From Theorem 13, the upper bound can be achieved by a suitably designed ϵ-stealthy attack. Thus, Fig. 2 represents a fundamental limitation for the performance degradation that can be induced by any ϵ-stealthy attack. Observe that plot is approximately linear as ϵ becomes large, as predicted by Corollary 9.

Example 2 Consider the system (A, B, C)

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & 0\\ 1 & -3 & 0 & 0 & 0\\ 0 & 0 & -2 & 0 & 0\\ 0 & 0 & 0 & -1 & 0\\ 0 & 0 & 0 & 0 & 3\end{bmatrix}, \quad B = \begin{bmatrix} 2 & 0\\ 1 & 0\\ 0 & 1\\ 0 & 1\\ 1 & 1\end{bmatrix},$$

$$C = \begin{bmatrix} 1 & -1 & 2 & 0 & 0\\ -1 & 2 & 0 & 3 & 0\\ 2 & 1 & 0 & 0 & 4\end{bmatrix},$$

which fails to be right invertible. Let $\Sigma_w = 0.5I$ and $\Sigma_v = I$. In Fig. 3, we plot the upper bound for the value of $\tilde{P}_W$ that an ϵ-stealthy attacker can induce, as calculated using Theorem 7. The value of $\tilde{P}_W$ achieved by the heuristic attack $\mathcal{A}_2$ is also plotted. Although the bound is fairly tight as compared to the performance degradation achieved by the heuristic attack; nonetheless, there remains a gap between the two plots.

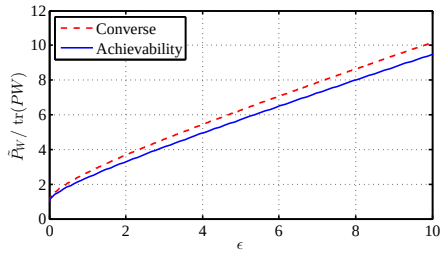

Fig. 3. The converse and achievability for the right non-invertible system with the weight matrix $W = C^T \Sigma_z^{-1} C$. The converse is obtained from (11) and the achievability is the weighted MSE $\tilde{P}_W$ induced by the heuristic algorithm $\mathcal{A}_2$.

6 Conclusion

This work characterizes fundamental limitations and achievability results for performance degradation induced by an attacker in a stochastic control system. The attacker is assumed to know the system parameters and noise statistics, and is able to hijack and replace the nominal control input. We propose a notion of ϵ-stealthiness to quantify the difficulty of detecting an attack from the measurements, and we characterize the largest degradation of Kalman filtering induced by an ϵ-stealthy attack. For right invertible systems, our study reveals that the nominal control input is the only critical piece of information to induce the largest performance degradation. For systems that are not right invertible, we provide an achievability result that lower bounds the performance degradation that an optimal ϵ-stealthy attack can achieve.

A Proof of Theorem 1

The first statement follows directly from the Neyman-Pearson Lemma Poor (1998).

For the second statement, we apply the Chernoff-Stein Lemma for ergodic measurements (see Polyanskiy & Wu (2012–2013)) that states that for any given attack sequence $\tilde{u}_1^\infty$, for a given $0 < 1 - p_k^D \leq \delta$ where $0 < \delta < 1$, the best achievable decay exponent of p_k^F is given by $\lim_{k\to\infty} \frac{1}{k} D\big(\tilde{y}_1^k \big\| y_1^k\big)$. For this attack sequence and with any detector, we obtain

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

Thus, by Definition 1, the attack is ϵ-stealthy.

Finally, the proof for the third statement follows by contradiction. Assume that (5) does not hold and there exists an ϵ-stealthy attack $\tilde{u}_1^\infty$ such that $\limsup_{k\to\infty} \frac{1}{k} D\big(\tilde{y}_1^k \big\| y_1^k\big) > \epsilon$. Suppose that the detector employs the standard log-likelihood ratio test with threshold λ_k at every time $k+1$. Thus, the test is $L_k(\eta_1^k) \underset{H_1}{\overset{H_0}{\lessgtr}} \lambda_k$, where $L_k(\eta_1^k) = \log \frac{f_{\tilde{y}_1^k}(\eta_1^k)}{f_{y_1^k}(\eta_1^k)}$ is the log-likelihood ratio and $\eta_1^k = y_1^k$ (resp. $\eta_1^k = \tilde{y}_1^k$) if H_0 (resp. H_1) is true. Define the conditional cumulant generating function for the log-likelihood ratio to be $g_{k|0}(s) = \log \mathbb{E}\big[e^{sL_k} \big| H_0\big]$ and $g_{k|1}(s) = \log \mathbb{E}\big[e^{sL_k} \big| H_1\big]$. Note that $g_{k|0}(s) = g_{k|1}(s-1)$. Let λ_k be chosen to ensure that $0 < 1 - p_k^D \leq \delta$ for every $k > 0$ (notice that such λ_k always exists, because p_k^D increases to one as λ_k decreases to zero). Then, for any $s_k > 0$, Chernoff's inequality yields

$$\begin{aligned} p_k^F &= \mathbb{P}[L_k \geq \lambda_k | H_0] \leq e^{-s_k\lambda_k + g_{k|0}(s_k)} \\ \Rightarrow -\log p_k^F &\geq s_k\lambda_k - g_{k|0}(s_k) \\ &\geq s_k\lambda_k - g_{k|1}(s_k - 1) \\ &= s_k\lambda_k - \log \mathbb{E}\big[e^{(s_k-1)L_k} \big| H_1\big]. \end{aligned}$$

Now, by applying Jensen's inequality twice we obtain

$$\begin{aligned} -\log p_k^F &\geq s_k\lambda_k + \log \mathbb{E}\big[e^{-(s_k-1)L_k} \big| H_1\big] \\ &\geq s_k\lambda_k + \mathbb{E}[-(s_k-1)L_k | H_1]. \end{aligned}$$

Finally, using $\mathbb{E}[L_k | H_1] = D\big(\tilde{y}_1^k \big\| y_1^k\big)$ implies

$$-\log p_k^F \geq D\big(\tilde{y}_1^k \big\| y_1^k\big) + s_k\Big(\lambda_k - D\big(\tilde{y}_1^k \big\| y_1^k\big)\Big). \quad \text{(A.1)}$$

Now, for any time index k such that $\frac{1}{k} D\big(\tilde{y}_1^k \big\| y_1^k\big) > \epsilon$, let

$$s_k = \frac{D\big(\tilde{y}_1^k \big\| y_1^k\big) - k\epsilon}{2\Big|D\big(\tilde{y}_1^k \big\| y_1^k\big) - \lambda_k\Big|}. \quad \text{(A.2)}$$

Using (A.1), (A.2) and $\limsup_{k\to\infty} \frac{1}{k} D\big(\tilde{y}_1^k \big\| y_1^k\big) > \epsilon$, we obtain $\limsup_{k\to\infty} -\frac{1}{k} \log p_k^F > \epsilon$, which contradicts the definition of ϵ-stealthiness. Hence, the attack cannot be stealthy, and the condition stated in (5) must be true.

B Proof of Lemma 4

By definition, we can write Kullback-Leibler divergence

$$\begin{aligned} D\big(\tilde{z}_1^k \big\| z_1^k\big) &= \int_{-\infty}^{\infty} f_{\tilde{z}_1^k}(t_1^k) \log f_{\tilde{z}_1^k}(t_1^k) dt_1^k \\ &\quad - \int_{-\infty}^{\infty} f_{\tilde{z}_1^k}(t_1^k) \log f_{z_1^k}(t_1^k) dt_1^k \\ &= -h\big(\tilde{z}_1^k\big) - \int_{-\infty}^{\infty} f_{\tilde{z}_1^k}(t_1^k) \log f_{z_1^k}(t_1^k) dt_1^k. \end{aligned}$$

Now, z_1^k is the innovation sequence without any attack and is thus an independent and identically distributed

sequence of Gaussian random variables with mean 0 and covariance Σ_z. Plugging into the above equation yields

$$D\big(\tilde{z}_1^k\big\|z_1^k\big) = -h\big(\tilde{z}_1^k\big) + \frac{k}{2}\log\Big((2\pi)^{N_y}\det(\Sigma_z)\Big) + \frac{1}{2}\sum_{n=1}^{k}\operatorname{tr}\big(\mathbb{E}\big[\tilde{z}_n\tilde{z}_n^T\big]\Sigma_z^{-1}\big),$$

which we can rewrite as

$$\frac{1}{2k}\sum_{n=1}^{k}\operatorname{tr}\big(\mathbb{E}\big[\tilde{z}_n\tilde{z}_n^T\big]\Sigma_z^{-1}\big) = \frac{1}{k}D\big(\tilde{z}_1^k\big\|z_1^k\big) - \frac{1}{2}\log\Big((2\pi)^{N_y}\det(\Sigma_z)\Big) + \frac{1}{k}h\big(\tilde{z}_1^k\big). \quad \text{(B.1)}$$

We can upper-bound the right hand side by first using the sub-additivity property of differential entropy (Cover & Thomas 2006, Corollary 8.6.1), and then further bounding the entropy $h(\tilde{z}_n)$ using the maximum differential entropy lemma (Gamal & Kim 2011, Section 2.2) for multivariate random variables. Thus, we obtain

$$\begin{aligned}
&\frac{1}{2k}\sum_{n=1}^{k}\operatorname{tr}\big(\mathbb{E}\big[\tilde{z}_n\tilde{z}_n^T\big]\Sigma_z^{-1}\big)\\
&\le \frac{1}{k}D\big(\tilde{z}_1^k\big\|z_1^k\big) - \frac{1}{2}\log\Big((2\pi)^{N_y}\det(\Sigma_z)\Big) + \frac{1}{k}\sum_{n=1}^{k}h(\tilde{z}_n)\\
&\le \frac{1}{k}D\big(\tilde{z}_1^k\big\|z_1^k\big) - \frac{1}{2}\log\Big((2\pi)^{N_y}\det(\Sigma_z)\Big) + \frac{1}{k}\sum_{n=1}^{k}\frac{1}{2}\log\big((2\pi e)^{N_y}\det(\mathbb{E}[\tilde{z}_n\tilde{z}_n^T])\big),
\end{aligned}$$

with equality if the sequence $\tilde{z}_1^k$ is an independent sequence of random variables with each random variable $\tilde{z}_n$ as Gaussian distributed with mean zero for all n. Straight-forward algebraic manipulation yields

$$\begin{aligned}
&\frac{1}{2k}\sum_{n=1}^{k}\operatorname{tr}\big(\mathbb{E}\big[\tilde{z}_n\tilde{z}_n^T\big]\Sigma_z^{-1}\big)\\
&\le \frac{1}{k}D\big(\tilde{z}_1^k\big\|z_1^k\big) - \frac{1}{2}\log\Big((2\pi)^{N_y}\Big) - \frac{1}{2}\log\Big(\det(\Sigma_z)\Big) + \frac{1}{k}\sum_{n=1}^{k}\frac{1}{2}\log\big((2\pi e)^{N_y}\big) + \frac{1}{k}\sum_{n=1}^{k}\frac{1}{2}\log\big(\det(\mathbb{E}[\tilde{z}_n\tilde{z}_n^T])\big)\\
&\le \frac{1}{k}D\big(\tilde{z}_1^k\big\|z_1^k\big) - \frac{1}{2}\log\Big((2\pi)^{N_y}\Big) + \frac{1}{k}\sum_{n=1}^{k}\frac{1}{2}\log\big((2\pi e)^{N_y}\big) + \frac{1}{k}\sum_{n=1}^{k}\frac{1}{2}\log\big(\det(\mathbb{E}[\tilde{z}_n\tilde{z}_n^T])\big) - \frac{1}{2}\log\Big(\det(\Sigma_z)\Big)\\
&= \frac{1}{k}D\big(\tilde{z}_1^k\big\|z_1^k\big) + \frac{N_y}{2} + \frac{1}{k}\sum_{n=1}^{k}\frac{1}{2}\log\big(\det(\mathbb{E}[\tilde{z}_n\tilde{z}_n^T])(\det\Sigma_z^{-1}\big).
\end{aligned}$$

We can further bound

$$\begin{aligned}
\det(\mathbb{E}[\tilde{z}_n\tilde{z}_n^T])(\det(\Sigma_z))^{-1} &= \det(\mathbb{E}[\tilde{z}_n\tilde{z}_n^T]\Sigma_z^{-1})\\
&\le \left(\frac{1}{N_y}\operatorname{tr}(\mathbb{E}[\tilde{z}_n\tilde{z}_n^T]\Sigma_z^{-1})\right)^{N_y},\\
\Rightarrow \frac{1}{2k}\sum_{n=1}^{k}\operatorname{tr}\big(\mathbb{E}\big[\tilde{z}_n\tilde{z}_n^T\big]\Sigma_z^{-1}\big) &\le \frac{1}{k}D\big(\tilde{z}_1^k\big\|z_1^k\big) + \frac{N_y}{2}\\
&\quad + \frac{N_y}{2k}\sum_{n=1}^{k}\log\Big(\frac{1}{N_y}\operatorname{tr}\big(\mathbb{E}[\tilde{z}_n\tilde{z}_n^T]\Sigma_z^{-1}\big)\Big),
\end{aligned}$$

with equality if the matrix $\mathbb{E}[\tilde{z}_n\tilde{z}_n^T]$ is a scalar multiplication of Σ_z for all n Finally, using the Arithmetic Mean and Geometric Mean (AM-GM) inequality yields the desired result (7). For the AM-GM inequality to hold with equality we need that $\operatorname{tr}(\mathbb{E}[\tilde{z}_n\tilde{z}_n^T]\Sigma_z^{-1})$ is constant for every n. Collecting all the above conditions for equality at various steps, (7) holds with equality if $\mathbb{E}\big[\tilde{z}_k\tilde{z}_k^T\big] = \alpha\Sigma_z$ for some scalar α.

References

Bai, C.-Z. & Gupta, V. (2014), On kalman filtering in the presence of a compromised sensor: Fundamental performance bounds, *in* 'American Control Conference', Portland, OR, pp. 3029–3034.

Bai, C.-Z., Pasqualetti, F. & Gupta, V. (2015), Security in stochastic control systems: Fundamental limitations and performance bounds, *in* 'American Control Conference', Chicago, IL, USA, pp. 195–200.

Basile, G. & Marro, G. (1991), *Controlled and Conditioned Invariants in Linear System Theory*, Prentice Hall.

Bhattacharya, S. & Başar, T. (2013), Differential game-theoretic approach to a spatial jamming problem, *in* 'Advances in Dynamic Games', Springer, pp. 245–268.

Chen, Y., Kar, S. & Moura, J. M. F. (2016), 'Optimal attack strategies subject to detection constraints against cyber-physical systems', *arXiv preprint arXiv:1610.03370* .

Cover, T. M. & Thomas, J. A. (2006), *Elements of Information Theory*, 2nd edn, Wiley.

Cui, S., Han, Z., Kar, S., Kim, T. T., Poor, H. V. & Tajer, A. (2012), 'Coordinated data-injection attack and detection in the smart grid: A detailed look at enriching detection solutions', *Signal Processing Magazine, IEEE* **29**(5), 106–115.

Dan, G. & Sandberg, H. (2010), Stealth attacks and protection schemes for state estimators in power systems, *in* 'IEEE Int. Conf. on Smart Grid Communications', Gaithersburg, MD, USA, pp. 214–219.

Farwell, J. P. & Rohozinski, R. (2011), 'Stuxnet and the future of cyber war', *Survival* **53**(1), 23–40.

Fawzi, H., Tabuada, P. & Diggavi, S. (2014), 'Secure estimation and control for cyber-physical systems under

adversarial attacks', *IEEE Transactions on Automatic Control* **59**(6), 1454–1467.

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

Gamal, A. E. & Kim, Y.-H. (2011), *Network information theory*, Cambridge University Press.

Giani, A., Bitar, E., Garcia, M., McQueen, M., Khargonekar, P. & Poolla, K. (2011), Smart grid data integrity attacks: characterizations and countermeasures, *in* 'IEEE Int. Conf. on Smart Grid Communications', Brussels, Belgium, pp. 232–237.

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* . To appear.

Hamza, F., Tabuada, P. & Diggavi, S. (2011), Secure state-estimation for dynamical systems under active adversaries, *in* 'Allerton Conf. on Communications, Control and Computing', pp. 337–344.

Hespanha, J. P. (2009), *Linear systems theory*, Princeton university press.

Kosut, O., Jia, L., Thomas, R. J. & Tong, L. (2011), 'Malicious data attacks on the smart grid', *IEEE Transactions on Smart Grid* **2**(4), 645–658.

Kullback, S. (1997), *Information theory and statistics*, Courier Dover Publications.

Kung, E., Dey, S. & Shi, L. (2016), 'The performance and limitations of ϵ-stealthy attacks on higher order systems', *IEEE Transactions on Automatic Control* . To appear.

Kuvshinkova, S. (2003), 'SQL Slammer worm lessons learned for consideration by the electricity sector', *North American Electric Reliability Council* .

Kwon, C., Liu, W. & Hwang, I. (2013), Security analysis for cyber-physical systems against stealthy deception attacks, *in* 'American Control Conference', IEEE, Washington, DC, USA, pp. 3344–3349.

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

Liu, Y., Reiter, M. K. & Ning, P. (2009), False data injection attacks against state estimation in electric power grids, *in* 'ACM Conference on Computer and Communications Security', Chicago, IL, USA, pp. 21–32.

Maharjan, S., Zhu, Q., Zhang, Y., Gjessing, S. & Başar, T. (2013), 'Dependable demand response management in the smart grid: A stackelberg game approach.', *IEEE Transactions Smart Grid* **4**(1), 120–132.

Manshaei, M., Zhu, Q., Alpcan, T., Başar, T. & Hubaux, J.-P. (2011), 'Game theory meets network security and privacy', *ACM Computing Surveys* **45**(3), 1–39.

Mo, Y., Chabukswar, R. & Sinopoli, B. (2014), 'Detecting integrity attacks on scada systems', *IEEE Transactions on Control Systems Technology* **22**(4), 1396–1407.

Mo, Y. & Sinopoli, B. (2010), Secure control against replay attacks, *in* 'Allerton Conf. on Communications, Control and Computing', Monticello, IL, USA, pp. 911–918.

Mohsenian-Rad, A.-H. & Leon-Garcia, A. (2011), 'Distributed internet-based load altering attacks against smart power grids', *IEEE Transactions on Smart Grid* **2**(4), 667–674.

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

Pasqualetti, F., Dörfler, F. & Bullo, F. (2015), 'Control-theoretic methods for cyberphysical security: Geometric principles for optimal cross-layer resilient control systems', *IEEE Control Systems Magazine* **35**(1), 110–127.

Patton, R., Frank, P. & Clark, R. (1989), *Fault Diagnosis in Dynamic Systems: Theory and Applications*, Prentice Hall.

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

Poor, H. V. (1998), *An introduction to signal detection and estimation*, 2nd edn, Springer-Verlag, New York.

Richards, G. (2008), 'Hackers vs slackers', *Engineering & Technology* **3**(19), 40–43.

Slay, J. & Miller, M. (2007), 'Lessons learned from the Maroochy water breach', *Critical Infrastructure Protection* **253**, 73–82.

Smith, R. (2011), A decoupled feedback structure for covertly appropriating network control systems, *in* 'IFAC World Congress', Milan, Italy, pp. 90–95.

Teixeira, A., Amin, S., Sandberg, H., Johansson, K. H. & Sastry, S. (2010), Cyber security analysis of state estimators in electric power systems, *in* 'IEEE Conf. on Decision and Control', Atlanta, GA, USA, pp. 5991–5998.

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

Weerakkody, S., Sinopoli, B., Kar, S. & Datta, A. (2016), 'Information flow for security in control systems', *arXiv preprint arXiv:1603.05710* .

Zhang, R. & Venkitasubramaniam, P. (2016), Stealthy control signal attacks in vector lqg systems, *in* 'American Control Conference', Boston, MA, USA, pp. 1179–1184.

Zhu, M. & Martínez, S. (2011), Stackelberg-game analysis of correlated attacks in cyber-physical systems, *in* 'American Control Conference', San Francisco, CA, USA, pp. 4063–4068.

Zhu, Q., Tembine, H. & Başar, T. (2013), 'Hybrid learning in stochastic games and its application in network security', *Reinforcement Learning and Approximate Dynamic Programming for Feedback Control* pp. 303–329.