Chapter 3
Fundamental Stealthiness–Distortion Trade-Offs in Cyber-Physical Systems

Song Fang and Quanyan Zhu

3.1 Introduction

Security issues such as the presence of malicious attacks could cause severe consequences in cyber-physical systems, which are safety-critical in most cases since they are interacting with the physical world. In the trend that cyber-physical systems are becoming more and more prevalent nowadays, it is also increasingly critical to be fully aware of such systems' performance limits (Fang et al. 2017), e.g., in terms of performance degradation, after taking the security issues into consideration. Accordingly, in this chapter, we focus on analyzing the fundamental limits of resilience in cyber-physical systems, including open-loop dynamical systems and (closed-loop) feedback control systems. More specifically, we examine the fundamental trade-offs between the systems' performance degradation that can be brought about by a malicious attack and the possibility of it being detected, of which the former is oftentimes measured by the mean squared-error distortion, whereas the latter is fundamentally determined by the Kullback–Leibler (KL) divergence.

The KL divergence was proposed in Kullback and Leibler (1951) (see also Kullback (1997)), and ever since it has been employed in various research areas, including, e.g., information theory (Cover and Thomas 2006), signal processing (Kay 2020), statistics (Pardo 2006), control and estimation theory (Lindquist and Picci 2015), system identification (Stoorvogel and Van Schuppen 1996), and machine learning (Goodfellow et al. 2016). Particularly, in statistical detection theory (Poor 2013), KL divergence provides the optimal exponent in probability of error for binary hypotheses testing problems as a result of the Chernoff–Stein lemma (Cover and Thomas 2006). Accordingly, in the context of determining whether an attack signal is present

S. Fang (✉) · Q. Zhu
New York University, 370 Jay Street, Brooklyn, New York 11201, USA
e-mail: song.fang@nyu.edu

Q. Zhu
e-mail: quanyan.zhu@nyu.edu

M. Abbaszadeh and A. Zemouche (eds.), *Security and Resilience in Cyber-Physical Systems*, https://doi.org/10.1007/978-3-030-97166-3_3

or not in security problems, the KL divergence has also been employed as a measure of stealthiness for attacks (see detailed discussions in, e.g., Bai et al. (2017a, b)).

In the context of dynamical and control system security (see, e.g., Poovendran et al. (2012), Johansson et al. (2014), Sandberg et al. (2015), Cheng et al. (2017), Giraldo et al. (2018), Weerakkody et al. (2019), Dibaji et al. (2019), Chong et al. (2019) and the references therein), particularly in dynamical and control systems under injection attacks, fundamental stealthiness–distortion trade-offs (with the mean squared-error as the distortion measure and the KL divergence as the stealthiness measure) have been investigated for feedback control systems (see, e.g., Zhang and Venkitasubramaniam (2017), Bai et al. (2017b)) as well as state estimation systems (see, e.g., Bai et al. (2017a), Kung et al. (2016), Guo et al. (2018)). Generally speaking, the problem considered is: Given a constraint (upper bound) on the level of stealthiness, what is the maximum degree of distortion (for control or for estimation) that can be caused by the attacker? This is dual to the following question: Given a least requirement (lower bound) on the degree of distortion, what is the maximum level of stealthiness that can be achieved by the attacker? Answers to these questions can not only capture the fundamental trade-offs between stealthiness and distortion but also characterize what the worst-case attacks are.

In this chapter, unlike the aforementioned works in Bai et al. (2017a, b), Kung et al. (2016), Zhang and Venkitasubramaniam (2017), Guo et al. (2018), we adopt an alternative approach to this stealthiness–distortion trade-off problem using power spectral analysis. The scenarios we consider include linear Gaussian open-loop dynamical systems and (closed-loop) feedback control systems. By using the power spectral approach, we obtain explicit formulas that characterize analytically the stealthiness–distortion trade-offs as well as the properties of the worst-case attacks. It turns out that the worst-case attacks are stationary colored Gaussian attacks with power spectra that are shaped specifically according to the transfer functions of the systems and the power spectra of the system outputs, the knowledge of which is all that the attacker needs to have access to in order to carry out the worst-case attacks. In other words, the attacker only needs to know the input–output behaviors of the systems, whereas it is not necessary to know their state-space models.

The remainder of the chapter is organized as follows. Section 3.2 provides the technical preliminaries. Section 3.3 is divided into two subsections, focusing on open-loop dynamical systems and feedback control systems, respectively. Section 3.4 presents numerical examples. Concluding remarks are given in Sect. 3.5.

More specifically, Theorem 3.1, as the first main result, characterizes explicitly the stealthiness–distortion trade-off and the worst-case attack in linear Gaussian open-loop dynamical systems. Equivalently, Corollary 3.1 considers the dual problem to that of Theorem 3.1. On the other hand, Theorem 3.2, together with Corollary 3.2 (in a dual manner), provides analytical expressions for the stealthiness–distortion trade-off and the worst-case attack in linear Gaussian feedback control systems. In addition, the preliminary results on the implications in control design, as presented in the Conclusion, indicate how the explicit stealthiness–distortion trade-off formula for feedback control systems can be employed to render the controller design explicit and intuitive.

Note that this chapter is based upon (Fang and Zhu 2021), which, however, only discusses the case of open-loop dynamical systems. Meanwhile, in this chapter, we also consider (closed-loop) feedback control systems. Note also that the results presented in this book chapter are applicable to discrete-time systems.

Notation: Throughout the chapter, we consider zero-mean real-valued continuous random variables and random vectors, as well as discrete-time stochastic processes. We represent random variables and random vectors using boldface letters, e.g., $\mathbf{x}$, while the probability density function of $\mathbf{x}$ is denoted as $p_{\mathbf{x}}$. In addition, $\mathbf{x}_{0,\ldots,k}$ will be employed to denote the sequence $\mathbf{x}_0, \ldots, \mathbf{x}_k$ or the random vector $\left[\mathbf{x}_0^{\mathrm{T}}, \ldots, \mathbf{x}_k^{\mathrm{T}}\right]^{\mathrm{T}}$, depending on the context. Note in particular that, for simplicity and with abuse of notations, we utilize $\mathbf{x} \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^m$ to indicate that $\mathbf{x}$ is a real-valued random variable and that $\mathbf{x}$ is a real-valued m-dimensional random vector, respectively.

3.2 Preliminaries

A stochastic process $\{\mathbf{x}_k\}, \mathbf{x}_k \in \mathbb{R}$ is said to be stationary if $R_{\mathbf{x}}(i,k) := \mathbb{E}\left[\mathbf{x}_i \mathbf{x}_{i+k}\right]$ depends only on k, and can thus be denoted as $R_{\mathbf{x}}(k)$ for simplicity. The power spectrum of a stationary process $\{\mathbf{x}_k\}, \mathbf{x}_k \in \mathbb{R}$ is defined as

$$S_{\mathbf{x}}(\omega) := \sum_{k=-\infty}^{\infty} R_{\mathbf{x}}(k) \, \mathrm{e}^{-\mathrm{j}\omega k}.$$

Moreover, the variance of $\{\mathbf{x}_k\}$ is given by

$$\sigma_{\mathbf{x}}^2 = \mathbb{E}\left[\mathbf{x}_k^2\right] = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_{\mathbf{x}}(\omega) \, \mathrm{d}\omega.$$

The KL divergence (see, e.g., Kullback and Leibler (1951)) is defined as follows.

Definition 3.1 Consider random vectors $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^m$ with probability densities $p_{\mathbf{x}}(\mathbf{u})$ and $p_{\mathbf{y}}(\mathbf{u})$, respectively. The KL divergence from distribution $p_{\mathbf{x}}$ to distribution $p_{\mathbf{y}}$ is defined as

$$\mathrm{KL}\left(p_{\mathbf{y}} \| p_{\mathbf{x}}\right) := \int p_{\mathbf{y}}(\mathbf{u}) \ln \frac{p_{\mathbf{y}}(\mathbf{u})}{p_{\mathbf{x}}(\mathbf{u})} \mathrm{d}\mathbf{u}.$$

The next lemma (see, e.g., Kay (2020)) provides an explicit expression of KL divergence in terms of covariance matrices for Gaussian random vectors; note that herein and in the sequel, all random variables and random vectors are assumed to be zero mean.

Lemma 3.1 *Consider Gaussian random vectors* $\mathbf{x} \in \mathbb{R}^m$ *and* $\mathbf{y} \in \mathbb{R}^m$ *with covariance matrices* $\Sigma_{\mathbf{x}}$ *and* $\Sigma_{\mathbf{y}}$*, respectively. The KL divergence from distribution* $p_{\mathbf{x}}$ *to distribution* $p_{\mathbf{y}}$ *is given by*

$$\mathrm{KL}\left(p_{\mathbf{y}} \| p_{\mathbf{x}}\right)=\frac{1}{2}\left[tr\left(\Sigma_{\mathbf{y}} \Sigma_{\mathbf{x}}^{-1}\right)-\ln \det\left(\Sigma_{\mathbf{y}} \Sigma_{\mathbf{x}}^{-1}\right)-m\right].$$

It is clear that in the scalar case (when $m=1$), Lemma 3.1 reduces to the following formula for Gaussian random variables:

$$\mathrm{KL}\left(p_{\mathbf{y}} \| p_{\mathbf{x}}\right)=\frac{1}{2}\left[\frac{\sigma_{\mathbf{y}}^{2}}{\sigma_{\mathbf{x}}^{2}}-\ln\left(\frac{\sigma_{\mathbf{y}}^{2}}{\sigma_{\mathbf{x}}^{2}}\right)-1\right].$$

The KL divergence rate (see, e.g., Lindquist and Picci (2015)) is defined as follows.

Definition 3.2 Consider stochastic processes $\{\mathbf{x}_k\}$, $\mathbf{x}_k \in \mathbb{R}^m$ and $\{\mathbf{y}_k\}$, $\mathbf{y}_k \in \mathbb{R}^m$ with densities $p_{\{\mathbf{x_k}\}}$ and $p_{\{\mathbf{y_k}\}}$, respectively; note that $p_{\{\mathbf{x_k}\}}$ and $p_{\{\mathbf{y_k}\}}$ will be denoted by $p_{\mathbf{x}}$ and $p_{\mathbf{y}}$ for simplicity in the sequel. Then, the KL divergence rate from distribution $p_{\mathbf{x}}$ to distribution $p_{\mathbf{y}}$ is defined as

$$\mathrm{KL}_{\infty}\left(p_{\mathbf{y}} \| p_{\mathbf{x}}\right):=\limsup_{k\rightarrow\infty}\frac{\mathrm{KL}\left(p_{\mathbf{y}_{0,\ldots,k}} \| p_{\mathbf{x}_{0,\ldots,k}}\right)}{k+1}.$$

The next lemma (see, e.g., Lindquist and Picci (2015)) provides an explicit expression of KL divergence rate in terms of power spectra for stationary Gaussian processes.

Lemma 3.2 *Consider stationary Gaussian processes* $\{\mathbf{x}_k\}$, $\mathbf{x}_k \in \mathbb{R}$ *and* $\{\mathbf{y}_k\}$, $\mathbf{y}_k \in \mathbb{R}$ *with densities* $p_{\mathbf{x}}$ *and* $p_{\mathbf{y}}$ *as well as power spectra* $S_{\mathbf{x}}(\omega)$ *and* $S_{\mathbf{y}}(\omega)$*, respectively. Suppose that* $S_{\mathbf{y}}(\omega)/S_{\mathbf{x}}(\omega)$ *is bounded (see Lindquist and Picci (2015) for details). Then, the KL divergence rate from distribution* $p_{\mathbf{x}}$ *to distribution* $p_{\mathbf{y}}$ *is given by*

$$\mathrm{KL}_{\infty}\left(p_{\mathbf{y}} \| p_{\mathbf{x}}\right)=\frac{1}{2\pi}\int_{0}^{2\pi}\frac{1}{2}\left\{\frac{S_{\mathbf{y}}(\omega)}{S_{\mathbf{x}}(\omega)}-\ln\left[\frac{S_{\mathbf{y}}(\omega)}{S_{\mathbf{x}}(\omega)}\right]-1\right\}\mathrm{d}\omega. \tag{3.1}$$

3.3 Stealthiness–Distortion Trade-Offs and Worst-Case Attacks

In this section, we analyze the fundamental stealthiness–distortion trade-offs of linear Gaussian open-loop dynamical systems and (closed-loop) feedback control systems under data injection attacks, whereas the KL divergence is employed as the stealthiness measure. Consider the scenario where attacker can modify the system input, and consequently, the system state and system output will then all be changed. From the attacker's point of view, the desired outcome is that the change in system state (as measured by state distortion) is large, while the change in system output (as measured by output stealthiness) is relatively small, so as to make the possibility of being detected low. Meanwhile fundamental trade-offs in general exist between state distortion and output stealthiness, since the system's state and output are correlated. In other words, increase in state distortion may inevitably lead to decrease in output stealthiness, i.e., increase in the possibility of being detected. How to capture such trade-offs? And what is the worst-case attack that can cause the maximum distortion given a certain stealthiness level, or vice versa? The answers are provided subsequently in terms of power spectral analysis.

3.3.1 Open-Loop Dynamical Systems

In this subsection, we focus on open-loop dynamical systems. Specifically, consider the scalar dynamical system depicted in Fig. 3.1 with state-space model given by

$$\begin{cases} \mathbf{x}_{k+1} = a\mathbf{x}_k + b\mathbf{u}_k + \mathbf{w}_k, \\ \mathbf{y}_k = c\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

where $\mathbf{x}_k \in \mathbb{R}$ is the system state, $\mathbf{u}_k \in \mathbb{R}$ is the system input, $\mathbf{y}_k \in \mathbb{R}$ is the system output, $\mathbf{w}_k \in \mathbb{R}$ is the process noise, and $\mathbf{v}_k \in \mathbb{R}$ is the measurement noise. The system parameters are $a \in \mathbb{R}$, $b \in \mathbb{R}$, and $c \in \mathbb{R}$; we further assume that $|a| < 1$ and $b, c \neq 0$, i.e., the system is stable, controllable, and observable. Accordingly, the transfer function of the system is given by

$$P(z) = \frac{bc}{z-a}. \tag{3.2}$$

(It is clear that $P(z)$ is minimum phase.) Suppose that $\{\mathbf{w}_k\}$ and $\{\mathbf{v}_k\}$ are stationary white Gaussian with variances $\sigma_{\mathbf{w}}^2$ and $\sigma_{\mathbf{v}}^2$, respectively. Furthermore, $\{\mathbf{w}_k\}$, $\{\mathbf{v}_k\}$, and $\mathbf{x}_0$ are assumed to be mutually independent. Assume also that $\{\mathbf{u}_k\}$ is stationary with power spectrum $S_{\mathbf{u}}(\omega)$. As such, $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$ are both stationary, and denote their power spectra by $S_{\mathbf{x}}(\omega)$ and $S_{\mathbf{y}}(\omega)$, respectively.

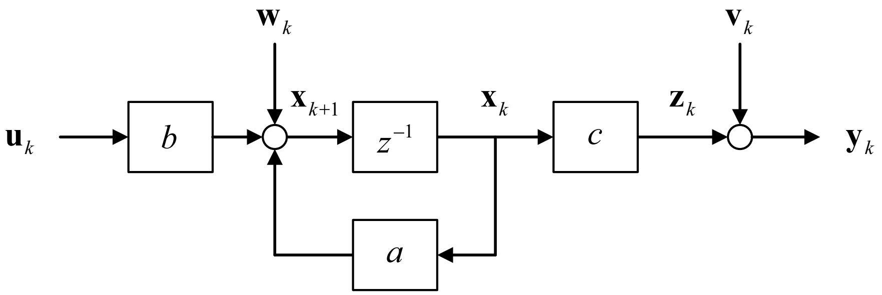

Fig. 3.1 A dynamical system

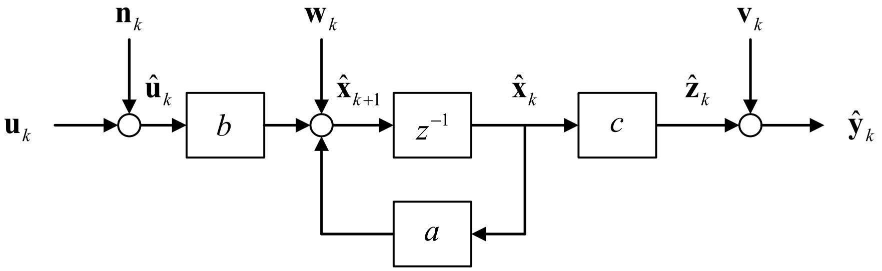

Fig. 3.2 A dynamical system under injection attack

Consider then the scenario that an attack signal $\{\mathbf{n}_k\}$, $\mathbf{n}_k \in \mathbb{R}$, is to be added to the input of the system $\{\mathbf{u}_k\}$ to deviate the system state, while aiming to be stealthy in the system output; see the depiction in Fig. 3.2. In addition, denote the true plant input under attack as $\{\widehat{\mathbf{u}}_k\}$, where

$$\widehat{\mathbf{u}}_k = \mathbf{u}_k + \mathbf{n}_k, \tag{3.3}$$

whereas the system under attack $\{\mathbf{n}_k\}$ is given by

$$\begin{cases} \widehat{\mathbf{x}}_{k+1} = a\widehat{\mathbf{x}}_k + b\widehat{\mathbf{u}}_k + \mathbf{w}_k = a\widehat{\mathbf{x}}_k + b\mathbf{u}_k + b\mathbf{n}_k + \mathbf{w}_k, \\ \quad \widehat{\mathbf{y}}_k = c\widehat{\mathbf{x}}_k + \mathbf{v}_k. \end{cases} \tag{3.4}$$

Meanwhile, suppose that the attack signal $\{\mathbf{n}_k\}$ is independent of $\{\mathbf{u}_k\}$, $\{\mathbf{w}_k\}$, $\{\mathbf{v}_k\}$, and $\mathbf{x}_0$; consequently, $\{\mathbf{n}_k\}$ is independent of $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$ as well.

The following questions then naturally arise: What is the fundamental trade-off between the degree of distortion caused in the system state (as measured by the mean squared-error distortion $\mathbb{E}\left[(\widehat{\mathbf{x}}_k - \mathbf{x}_k)^2\right]$ between the original state $\{\mathbf{x}_k\}$ and the state under attack denoted as $\{\widehat{\mathbf{x}}_k\}$) and the level of stealthiness resulted in the system output (as measured by the KL divergence rate $\mathrm{KL}_\infty\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right)$ between the original output $\{\mathbf{y}_k\}$ and the output under attack denoted as $\{\widehat{\mathbf{y}}_k\}$)? More specifically, to achieve a certain degree of distortion in state, what is the maximum level of stealthiness that can be maintained by the attacker? And what is the worst-case attack in this sense? The following theorem, as the first main result of this chapter, answers the questions raised above.

Theorem 3.1 *Consider the dynamical system under injection attacks depicted in Fig. 3.2. Suppose that the attacker aims to design the attack signal* $\{\mathbf{n}_k\}$ *to satisfy the following attack goal in terms of state distortion:*

$$\mathbb{E}\left[(\widehat{\mathbf{x}}_k - \mathbf{x}_k)^2\right] \geq D. \tag{3.5}$$

Then, the minimum KL divergence rate between the original output and the attacked output is given by

$$\inf_{\mathbb{E}\left[(\widehat{\mathbf{x}}_k-\mathbf{x}_k)^2\right]\geq D} \mathrm{KL}_{\infty}\left(p_{\widehat{\mathbf{y}}}\| p_{\mathbf{y}}\right) = \frac{1}{2\pi}\int_{0}^{2\pi}\frac{1}{2}\left\{\frac{S_{\widehat{\mathbf{n}}}\left(\omega\right)}{S_{\mathbf{y}}\left(\omega\right)} - \ln\left[1+\frac{S_{\widehat{\mathbf{n}}}\left(\omega\right)}{S_{\mathbf{y}}\left(\omega\right)}\right]\right\}\mathrm{d}\omega, \quad (3.6)$$

where

$$S_{\widehat{\mathbf{n}}}\left(\omega\right) = \frac{\zeta S_{\mathbf{y}}^{2}\left(\omega\right)}{1-\zeta S_{\mathbf{y}}\left(\omega\right)}, \quad (3.7)$$

and $S_{\mathbf{y}}\left(\omega\right)$ is given by

$$S_{\mathbf{y}}\left(\omega\right) = \frac{b^2c^2}{\left|\mathrm{e}^{\mathrm{j}\omega}-a\right|^2}S_{\mathbf{u}}\left(\omega\right) + \frac{c^2}{\left|\mathrm{e}^{\mathrm{j}\omega}-a\right|^2}\sigma_{\mathbf{w}}^2 + \sigma_{\mathbf{v}}^2. \quad (3.8)$$

Herein, ζ is the unique constant that satisfies

$$\frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\zeta S_{\mathbf{y}}^{2}\left(\omega\right)}{1-\zeta S_{\mathbf{y}}\left(\omega\right)}\mathrm{d}\omega = c^2 D, \quad (3.9)$$

while

$$0 < \zeta < \min_{\omega}\frac{1}{S_{\mathbf{y}}\left(\omega\right)}. \quad (3.10)$$

Moreover, the worst-case (in the sense of achieving this minimum KL divergence rate) attack $\{\mathbf{n}_k\}$ is a stationary colored Gaussian process with power spectrum

$$S_{\mathbf{n}}\left(\omega\right) = \frac{\left|\mathrm{e}^{\mathrm{j}\omega}-a\right|^2}{b^2c^2}\frac{\zeta S_{\mathbf{y}}^{2}\left(\omega\right)}{1-\zeta S_{\mathbf{y}}\left(\omega\right)}. \quad (3.11)$$

Proof To begin with, it can be verified that the power spectrum of $\{\mathbf{y}_k\}$ is given by

$$\begin{aligned} S_{\mathbf{y}}\left(\omega\right) &= \left|P\left(\mathrm{e}^{\mathrm{j}\omega}\right)\right|^2 S_{\mathbf{u}}\left(\omega\right) + \frac{1}{b^2}\left|P\left(\mathrm{e}^{\mathrm{j}\omega}\right)\right|^2\sigma_{\mathbf{w}}^2 + \sigma_{\mathbf{v}}^2, \\ &= \frac{b^2c^2}{\left|\mathrm{e}^{\mathrm{j}\omega}-a\right|^2}S_{\mathbf{u}}\left(\omega\right) + \frac{c^2}{\left|\mathrm{e}^{\mathrm{j}\omega}-a\right|^2}\sigma_{\mathbf{w}}^2 + \sigma_{\mathbf{v}}^2. \end{aligned}$$

Note then that due to the property of additivity of linear systems, the system in Fig. 3.2 is equivalent to that of Fig. 3.3, where

$$\widehat{\mathbf{y}}_k = \mathbf{y}_k + \widehat{\mathbf{n}}_k,$$

and $\{\widehat{\mathbf{n}}_k\}$ is the output of the subsystem

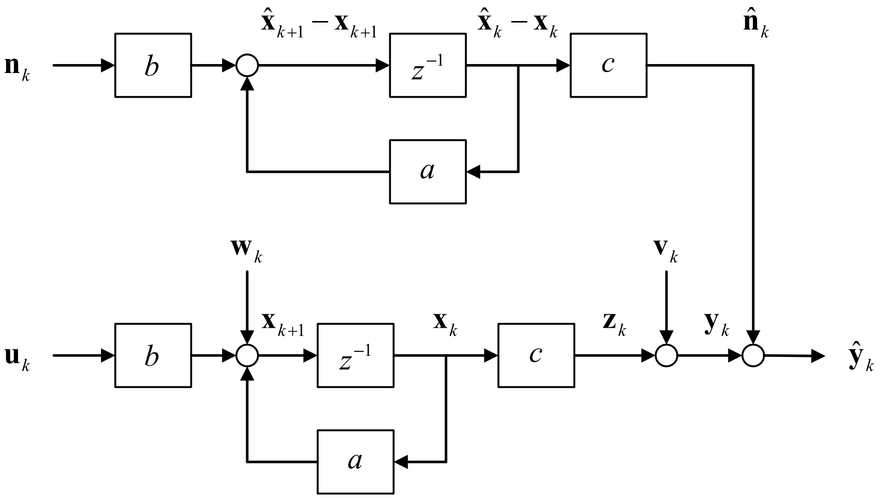

Fig. 3.3 A dynamical system under injection attack: equivalent system

$$\begin{cases} \widehat{\mathbf{x}}_{k+1} - \mathbf{x}_{k+1} = a\left(\widehat{\mathbf{x}}_k - \mathbf{x}_k\right) + b\mathbf{n}_k, \\ \widehat{\mathbf{n}}_k = c\left(\widehat{\mathbf{x}}_k - \mathbf{x}_k\right), \end{cases}$$

as depicted by the upper half of Fig. 3.3; note that in this subsystem, $(\widehat{\mathbf{x}}_k - \mathbf{x}_k) \in \mathbb{R}$ is the system state, $\mathbf{n}_k \in \mathbb{R}$ is the system input, and $\widehat{\mathbf{n}} \in \mathbb{R}$ is the system output. On the other hand, the distortion constraint

$$\mathbb{E}\left[(\widehat{\mathbf{x}}_k - \mathbf{x}_k)^2\right] \geq D$$

is then equivalent to being with a power constraint

$$\mathbb{E}\left[\widehat{\mathbf{n}}_k^2\right] \geq c^2 D,$$

since $\widehat{\mathbf{n}}_k = \widehat{\mathbf{y}}_k - \mathbf{y}_k$ and thus

$$\widehat{\mathbf{n}}_k^2 = (\mathbf{y}_k - \widehat{\mathbf{y}}_k)^2 = (c\mathbf{x}_k - c\widehat{\mathbf{x}}_k)^2 = c^2\left(\mathbf{x}_k - \widehat{\mathbf{x}}_k\right)^2.$$

Accordingly, the system in Fig. 3.3 may be viewed as a “virtual channel” modeled as

$$\widehat{\mathbf{y}}_k = \mathbf{y}_k + \widehat{\mathbf{n}}_k$$

with noise constraint

$$\mathbb{E}\left[\widehat{\mathbf{n}}_k^2\right] \geq c^2 D,$$

where $\{\mathbf{y}_k\}$ is the channel input, $\{\widehat{\mathbf{y}}_k\}$ is the channel output, and $\{\widehat{\mathbf{n}}_k\}$ is the channel noise. In addition, due to the fact that $\{\mathbf{n}_k\}$ is independent of $\{\mathbf{y}_k\}$, $\{\widehat{\mathbf{n}}_k\}$ is also independent of $\{\mathbf{y}_k\}$.

The approach we shall take herein, as developed in Cover and Thomas (2006), is to treat the multiple uses of a scalar channel (i.e., a scalar dynamic channel) equivalently as a single use of parallel channels (i.e., a set of parallel static channels). We consider first the case of a finite number of parallel static channels with

$$\widehat{\mathbf{y}} = \mathbf{y} + \widehat{\mathbf{n}},$$

where $\mathbf{y}, \widehat{\mathbf{y}}, \widehat{\mathbf{n}} \in \mathbb{R}^m$, and $\widehat{\mathbf{n}}$ is independent of $\mathbf{y}$. In addition, $\mathbf{y}$ is Gaussian with covariance $\Sigma_{\mathbf{y}}$, and the noise power constraint is given by

$$\operatorname{tr}\left(\Sigma_{\widehat{\mathbf{n}}}\right) = \mathbb{E}\left[\sum_{i=1}^{m} \widehat{\mathbf{n}}^2\left(i\right)\right] \geq c^2 D,$$

where $\widehat{\mathbf{n}}\left(i\right)$ denotes the i-th element of $\widehat{\mathbf{n}}$. In addition, according to Fang and Zhu (2020) (see Proposition 2 therein), we have

$$\operatorname{KL}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right) \geq \operatorname{KL}\left(p_{\widehat{\mathbf{y}}^{\mathrm{G}}} \| p_{\mathbf{y}}\right),$$

where $\widehat{\mathbf{y}}^{\mathrm{G}}$ denotes a Gaussian random vector with the same covariance as $\widehat{\mathbf{y}}$, and equality holds if $\widehat{\mathbf{y}}$ is Gaussian. Meanwhile, it is known from Lemma 3.1 that

$$\operatorname{KL}\left(p_{\widehat{\mathbf{y}}^{\mathrm{G}}} \| p_{\mathbf{y}}\right) = \frac{1}{2}\left[\operatorname{tr}\left(\Sigma_{\widehat{\mathbf{y}}}\Sigma_{\mathbf{y}}^{-1}\right) - \ln\det\left(\Sigma_{\widehat{\mathbf{y}}}\Sigma_{\mathbf{y}}^{-1}\right) - m\right].$$

On the other hand, since $\mathbf{y}$ and $\widehat{\mathbf{n}}$ are independent, we have

$$\Sigma_{\widehat{\mathbf{y}}} = \Sigma_{\widehat{\mathbf{n}}+\mathbf{y}} = \Sigma_{\widehat{\mathbf{n}}} + \Sigma_{\mathbf{y}}.$$

Consequently,

$$\operatorname{tr}\left(\Sigma_{\widehat{\mathbf{y}}}\Sigma_{\mathbf{y}}^{-1}\right) - \ln\det\left(\Sigma_{\widehat{\mathbf{y}}}\Sigma_{\mathbf{y}}^{-1}\right) = \operatorname{tr}\left[\left(\Sigma_{\widehat{\mathbf{n}}} + \Sigma_{\mathbf{y}}\right)\Sigma_{\mathbf{y}}^{-1}\right] - \ln\det\left[\left(\Sigma_{\widehat{\mathbf{n}}} + \Sigma_{\mathbf{y}}\right)\Sigma_{\mathbf{y}}^{-1}\right].$$

Denote the eigendecomposition of $\Sigma_{\mathbf{y}}$ by $U_{\mathbf{y}}\Lambda_{\mathbf{y}}U_{\mathbf{y}}^{\mathrm{T}}$, where

$$\Lambda_{\mathbf{y}} = \operatorname{diag}\left(\lambda_1, \ldots, \lambda_m\right).$$

Then,

$$\begin{aligned}
&\operatorname{tr}\left[\left(\Sigma_{\widehat{\mathbf{n}}}+\Sigma_{\mathrm{y}}\right)\Sigma_{\mathrm{y}}^{-1}\right]-\ln\det\left[\left(\Sigma_{\widehat{\mathbf{n}}}+\Sigma_{\mathrm{y}}\right)\Sigma_{\mathrm{y}}^{-1}\right]\\
&\quad=\operatorname{tr}\left[\left(\Sigma_{\widehat{\mathbf{n}}}+U_{\mathrm{y}}\Lambda_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\right)\left(U_{\mathrm{y}}\Lambda_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\right)^{-1}\right]-\ln\det\left[\left(\Sigma_{\widehat{\mathbf{n}}}+U_{\mathrm{y}}\Lambda_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\right)\left(U_{\mathrm{y}}\Lambda_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\right)^{-1}\right],\\
&\quad=\operatorname{tr}\left[\left(\Sigma_{\widehat{\mathbf{n}}}+U_{\mathrm{y}}\Lambda_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\right)U_{\mathrm{y}}\Lambda_{\mathrm{y}}^{-1}U_{\mathrm{y}}^{\mathrm{T}}\right]-\ln\det\left[\left(\Sigma_{\widehat{\mathbf{n}}}+U_{\mathrm{y}}\Lambda_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\right)U_{\mathrm{y}}\Lambda_{\mathrm{y}}^{-1}U_{\mathrm{y}}^{\mathrm{T}}\right],\\
&\quad=\operatorname{tr}\left[U_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\left(\Sigma_{\widehat{\mathbf{n}}}+U_{\mathrm{y}}\Lambda_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\right)U_{\mathrm{y}}\Lambda_{\mathrm{y}}^{-1}U_{\mathrm{y}}^{\mathrm{T}}\right]\\
&\qquad-\ln\det\left[U_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\left(\Sigma_{\widehat{\mathbf{n}}}+U_{\mathrm{y}}\Lambda_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\right)U_{\mathrm{y}}\Lambda_{\mathrm{y}}^{-1}U_{\mathrm{y}}^{\mathrm{T}}\right],\\
&\quad=\operatorname{tr}\left\{U_{\mathrm{y}}\left[U_{\mathrm{y}}^{\mathrm{T}}\left(\Sigma_{\widehat{\mathbf{n}}}+U_{\mathrm{y}}\Lambda_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\right)U_{\mathrm{y}}\Lambda_{\mathrm{y}}^{-1}\right]U_{\mathrm{y}}^{\mathrm{T}}\right\}\\
&\qquad-\ln\det\left\{U_{\mathrm{y}}\left[U_{\mathrm{y}}^{\mathrm{T}}\left(\Sigma_{\widehat{\mathbf{n}}}+U_{\mathrm{y}}\Lambda_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\right)U_{\mathrm{y}}\Lambda_{\mathrm{y}}^{-1}\right]U_{\mathrm{y}}^{\mathrm{T}}\right\},\\
&\quad=\operatorname{tr}\left[U_{\mathrm{y}}^{\mathrm{T}}\left(\Sigma_{\widehat{\mathbf{n}}}+U_{\mathrm{y}}\Lambda_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\right)U_{\mathrm{y}}\Lambda_{\mathrm{y}}^{-1}\right]-\ln\det\left[U_{\mathrm{y}}^{\mathrm{T}}\left(\Sigma_{\widehat{\mathbf{n}}}+U_{\mathrm{y}}\Lambda_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\right)U_{\mathrm{y}}\Lambda_{\mathrm{y}}^{-1}\right],\\
&\quad=\operatorname{tr}\left[\left(U_{\mathrm{y}}^{\mathrm{T}}\Sigma_{\widehat{\mathbf{n}}}U_{\mathrm{y}}+\Lambda_{\mathrm{y}}\right)\Lambda_{\mathrm{y}}^{-1}\right]-\ln\det\left[\left(U_{\mathrm{y}}^{\mathrm{T}}\Sigma_{\widehat{\mathbf{n}}}U_{\mathrm{y}}+\Lambda_{\mathrm{y}}\right)\Lambda_{\mathrm{y}}^{-1}\right],\\
&\quad=\operatorname{tr}\left[\left(\overline{\Sigma}_{\widehat{\mathbf{n}}}+\Lambda_{\mathrm{y}}\right)\Lambda_{\mathrm{y}}^{-1}\right]-\ln\det\left[\left(\overline{\Sigma}_{\widehat{\mathbf{n}}}+\Lambda_{\mathrm{y}}\right)\Lambda_{\mathrm{y}}^{-1}\right],
\end{aligned}$$

where $\overline{\Sigma}_{\widehat{\mathbf{n}}}=U_{\mathrm{y}}^{\mathrm{T}}\Sigma_{\widehat{\mathbf{n}}}U_{\mathrm{y}}$. Denoting the diagonal terms of $\overline{\Sigma}_{\widehat{\mathbf{n}}}$ by $\overline{\sigma}_{\widehat{\mathbf{n}}(i)}^{2}$, $i=1,\ldots,m$, it is known from (Fang and Zhu 2020) (see Proposition 4 therein) that

$$\begin{aligned}
&\operatorname{tr}\left[\left(\overline{\Sigma}_{\widehat{\mathbf{n}}}+\Lambda_{\mathrm{y}}\right)\Lambda_{\mathrm{y}}^{-1}\right]-\ln\det\left[\left(\overline{\Sigma}_{\widehat{\mathbf{n}}}+\Lambda_{\mathrm{y}}\right)\Lambda_{\mathrm{y}}^{-1}\right],\\
&\qquad\geq\sum_{i=1}^{m}\left[\frac{\overline{\sigma}_{\widehat{\mathbf{n}}(i)}^{2}+\lambda_i}{\lambda_i}\right]-\sum_{i=1}^{m}\ln\left[\frac{\overline{\sigma}_{\widehat{\mathbf{n}}(i)}^{2}+\lambda_i}{\lambda_i}\right],\\
&\qquad=\sum_{i=1}^{m}\left[1+\frac{\overline{\sigma}_{\widehat{\mathbf{n}}(i)}^{2}}{\lambda_i}\right]-\sum_{i=1}^{m}\ln\left[1+\frac{\overline{\sigma}_{\widehat{\mathbf{n}}(i)}^{2}}{\lambda_i}\right],
\end{aligned}$$

where equality holds if $\overline{\Sigma}_{\widehat{\mathbf{n}}}$ is diagonal. For simplicity, we denote

$$\overline{\Sigma}_{\widehat{\mathbf{n}}}=\operatorname{diag}\left(\overline{\sigma}_{\widehat{\mathbf{n}}(1)}^{2},\ldots,\overline{\sigma}_{\widehat{\mathbf{n}}(m)}^{2}\right)=\operatorname{diag}\left(\widehat{N}_1,\ldots,\widehat{N}_m\right)$$

when $\overline{\Sigma}_{\widehat{\mathbf{n}}}$ is diagonal. Then, the problem reduces to that of choosing $\widehat{N}_1,\ldots,\widehat{N}_m$ to minimize

$$\sum_{i=1}^{m}\left(1+\frac{\widehat{N}_i}{\lambda_i}\right)-\sum_{i=1}^{m}\ln\left(1+\frac{\widehat{N}_i}{\lambda_i}\right)$$

subject to the constraint that

$$\sum_{i=1}^{m}\widehat{N}_i=\operatorname{tr}\left(\overline{\Sigma}_{\widehat{\mathbf{n}}}\right)=\operatorname{tr}\left(U_{\mathrm{y}}^{\mathrm{T}}\Sigma_{\widehat{\mathbf{n}}}U_{\mathrm{y}}\right)=\operatorname{tr}\left(\Sigma_{\widehat{\mathbf{n}}}U_{\mathrm{y}}U_{\mathrm{y}}^{\mathrm{T}}\right)=\operatorname{tr}\left(\Sigma_{\widehat{\mathbf{n}}}\right)=mc^{2}D.$$

Define the Lagrange function by

$$\sum_{i=1}^{m}\left(1+\frac{\widehat{N}_i}{\lambda_i}\right)-\sum_{i=1}^{m}\ln\left(1+\frac{\widehat{N}_i}{\lambda_i}\right)+\eta\left(\sum_{i=1}^{m}\widehat{N}_i-\widehat{N}\right),$$

and differentiate it with respect to $\widehat{N}_i$, then we have

$$\frac{1}{\lambda_i}-\frac{1}{\widehat{N}_i+\lambda_i}+\eta=0,$$

or equivalently,

$$\widehat{N}_i=\frac{1}{\frac{1}{\lambda_i}+\eta}-\lambda_i=\frac{\lambda_i}{1+\eta\lambda_i}-\lambda_i=\frac{-\eta\lambda_i^2}{1+\eta\lambda_i},$$

where η satisfies

$$\sum_{i=1}^{m}\widehat{N}_i=\sum_{i=1}^{m}\frac{-\eta\lambda_i^2}{1+\eta\lambda_i}=mc^2D,$$

while

$$-\min_{i=0,\ldots,m}\frac{1}{\lambda_i}<\eta<0.$$

For simplicity, we denote $\zeta=-\eta$, and accordingly,

$$\widehat{N}_i=\frac{\zeta\lambda_i^2}{1-\zeta\lambda_i},$$

where ζ satisfies

$$\sum_{i=1}^{m}\widehat{N}_i=\sum_{i=1}^{m}\frac{\zeta\lambda_i^2}{1-\zeta\lambda_i}=mc^2D,$$

while

$$0<\zeta<\min_{i=0,\ldots,m}\frac{1}{\lambda_i}.$$

Correspondingly,

$$\begin{aligned}\inf_{p_{\widehat{\mathbf{n}}}} \mathrm{KL}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right) &= \frac{1}{2}\left[\sum_{i=1}^{m}\left(1+\frac{\widehat{N}_i}{\lambda_i}\right)-\sum_{i=1}^{m}\ln\left(1+\frac{\widehat{N}_i}{\lambda_i}\right)-m\right],\\ &= \sum_{i=1}^{m}\frac{1}{2}\left[\frac{\widehat{N}_i}{\lambda_i}-\ln\left(1+\frac{\widehat{N}_i}{\lambda_i}\right)\right].\end{aligned}$$

Consider now a scalar dynamic channel

$$\widehat{\mathbf{y}}_k = \mathbf{y}_k + \widehat{\mathbf{n}}_k,$$

where $\mathbf{y}_k, \widehat{\mathbf{n}}_k, \widehat{\mathbf{y}}_k \in \mathbb{R}$, while $\{\mathbf{y}_k\}$ and $\{\widehat{\mathbf{n}}_k\}$ are independent. In addition, $\{\mathbf{y}_k\}$ is stationary colored Gaussian with power spectrum $S_{\mathbf{y}}(\omega)$, whereas the noise power constraint is given by $\mathbb{E}\left[\widehat{\mathbf{n}}_k^2\right] \geq c^2 D$. We may then consider a block of consecutive uses from time 0 to k of this channel as $k+1$ channels in parallel Cover and Thomas (2006). Particularly, let the eigendecomposition of $\Sigma_{\mathbf{y}_{0,\ldots,k}}$ be given by

$$\Sigma_{\mathbf{y}_{0,\ldots,k}} = U_{\mathbf{y}_{0,\ldots,k}} \Lambda_{\mathbf{y}_{0,\ldots,k}} U_{\mathbf{y}_{0,\ldots,k}}^{\mathrm{T}},$$

where

$$\Lambda_{\mathbf{y}_{0,\ldots,k}} = \operatorname{diag}\left(\lambda_0, \ldots, \lambda_k\right).$$

Then, we have

$$\min_{p_{\widehat{\mathbf{n}}_{0,\ldots,k}}:\ \sum_{i=0}^{k}\mathbb{E}\left[\widehat{\mathbf{n}}_i^2\right]\geq(k+1)c^2D} \frac{\mathrm{KL}\left(p_{\widehat{\mathbf{y}}_{0,\ldots,k}} \| p_{\mathbf{y}_{0,\ldots,k}}\right)}{k+1} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{1}{2}\left[\frac{\widehat{N}_i}{\lambda_i}-\ln\left(1+\frac{\widehat{N}_i}{\lambda_i}\right)\right],$$

where

$$\widehat{N}_i = \frac{\zeta\lambda_i^2}{1-\zeta\lambda_i},\ i=0,\ldots,k.$$

Herein, ζ satisfies

$$\sum_{i=0}^{k}\widehat{N}_i = \sum_{i=0}^{k}\frac{\zeta\lambda_i^2}{1-\zeta\lambda_i} = (k+1)\,c^2D,$$

or equivalently,

$$\frac{1}{k+1}\sum_{i=0}^{k}\widehat{N}_i = \frac{1}{k+1}\left(\frac{\zeta\lambda_i^2}{1-\zeta\lambda_i}\right) = c^2D,$$

while

$$0 < \zeta < \min_{i=0,\ldots,k} \frac{1}{\lambda_i}.$$

In addition, since the processes $\{\mathbf{y}_k\}$, $\{\widehat{\mathbf{n}}_k\}$, and $\{\widehat{\mathbf{y}}_k\}$ are stationary, we have

$$\begin{aligned}
&\lim_{k\to\infty} \min_{p_{\widehat{\mathbf{n}}_{0,\ldots,k}}:\ \sum_{i=0}^{k}\mathbb{E}\left[\widehat{\mathbf{n}}_i^2\right]\geq (k+1)c^2 D} \frac{\mathrm{KL}\left(p_{\widehat{\mathbf{y}}_{0,\ldots,k}} \| p_{\mathbf{y}_{0,\ldots,k}}\right)}{k+1} \\
&= \inf_{\mathbb{E}\left[\widehat{\mathbf{n}}_k^2\right]\geq c^2 D} \lim_{k\to\infty} \frac{\mathrm{KL}\left(p_{\widehat{\mathbf{y}}_{0,\ldots,k}} \| p_{\mathbf{y}_{0,\ldots,k}}\right)}{k+1} = \inf_{\mathbb{E}\left[\widehat{\mathbf{n}}_k^2\right]\geq c^2 D} \limsup_{k\to\infty} \frac{\mathrm{KL}\left(p_{\widehat{\mathbf{y}}_{0,\ldots,k}} \| p_{\mathbf{y}_{0,\ldots,k}}\right)}{k+1} \\
&= \inf_{\mathbb{E}\left[\widehat{\mathbf{n}}_k^2\right]\geq c^2 D} \mathrm{KL}_{\infty}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right) = \inf_{\mathbb{E}\left[\left(\widehat{\mathbf{x}}_k-\mathbf{x}_k\right)^2\right]\geq D} \mathrm{KL}_{\infty}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right).
\end{aligned}$$

On the other hand, since the processes are stationary, the covariance matrices are Toeplitz (Grenander and Szegö 1958), and their eigenvalues approach their limits as $k \to \infty$. Moreover, the densities of eigenvalues on the real line tend to the power spectra of the processes (Gutiérrez-Gutiérrez and Crespo 2008; Lindquist and Picci 2015; Pinsker 1964). Accordingly,

$$\begin{aligned}
&\inf_{\mathbb{E}\left[\left(\widehat{\mathbf{x}}_k-\mathbf{x}_k\right)^2\right]\geq D} \mathrm{KL}_{\infty}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right) = \lim_{k\to\infty} \frac{1}{k+1} \sum_{i=0}^{k} \frac{1}{2}\left[\frac{\widehat{N}_i}{\lambda_i} - \ln\left(1+\frac{\widehat{N}_i}{\lambda_i}\right)\right], \\
&= \frac{1}{2\pi}\int_{0}^{2\pi} \frac{1}{2}\left\{\frac{S_{\widehat{\mathbf{n}}}(\omega)}{S_{\mathbf{y}}(\omega)} - \ln\left[1+\frac{S_{\widehat{\mathbf{n}}}(\omega)}{S_{\mathbf{y}}(\omega)}\right]\right\}\mathrm{d}\omega,
\end{aligned}$$

where

$$S_{\widehat{\mathbf{n}}}(\omega) = \frac{\zeta S_{\mathbf{y}}^2(\omega)}{1-\zeta S_{\mathbf{y}}(\omega)},$$

and ζ satisfies

$$\lim_{k\to\infty}\frac{1}{k+1}\sum_{i=0}^{k}\widehat{N}_i = \frac{1}{2\pi}\int_{-\pi}^{\pi} S_{\widehat{\mathbf{n}}}(\omega)\,\mathrm{d}\omega = \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\zeta S_{\mathbf{y}}^2(\omega)}{1-\zeta S_{\mathbf{y}}(\omega)}\mathrm{d}\omega = c^2 D,$$

while

$$0 < \zeta < \min_{\omega} \frac{1}{S_{\mathbf{y}}(\omega)}.$$

Lastly, note that

$$S_{\widehat{\mathbf{n}}}(\omega) = \left|P\left(\mathrm{e}^{\mathrm{j}\omega}\right)\right|^2 S_{\mathbf{n}}(\omega) = \frac{b^2c^2}{\left|\mathrm{e}^{\mathrm{j}\omega}-a\right|^2} S_{\mathbf{n}}(\omega),$$

and hence

$$S_{\mathbf{n}}(\omega) = \frac{\left|e^{j\omega} - a\right|^2}{b^2 c^2} S_{\widehat{\mathbf{n}}}(\omega) = \frac{\left|e^{j\omega} - a\right|^2}{b^2 c^2} \frac{\zeta S_{\mathbf{y}}^2(\omega)}{1 - \zeta S_{\mathbf{y}}(\omega)}.$$

This concludes the proof. ■

It is clear that $S_{\mathbf{n}}(\omega)$ may be rewritten as

$$S_{\mathbf{n}}(\omega) = \frac{1}{\left|P\left(e^{j\omega}\right)\right|^2} \frac{\zeta S_{\mathbf{y}}^2(\omega)}{1 - \zeta S_{\mathbf{y}}(\omega)}. \tag{3.12}$$

This means that the attacker only needs the knowledge of the power spectrum of the original system output $\{\mathbf{y}_k\}$ and the transfer function of the system (from $\{\mathbf{n}_k\}$ to $\{\widehat{\mathbf{y}}_k\}$), i.e., $P(z)$, in order to carry out this worst-case attack. It is worth mentioning that the power spectrum of $\{\mathbf{y}_k\}$ can be estimated based on its realizations (see, e.g., Stoica and Moses (2005)), while the transfer function of the system can be approximated by system identification (see, e.g., Ljung (1999)).

Note that it can be verified (Kay 2020) that the (minimum) output KL divergence rate $\mathrm{KL}_\infty\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right)$ increases strictly with the state distortion bound D. In other words, in order for the attacker to achieve larger distortion, the stealthiness level of the attack will inevitably decrease.

On the other hand, the dual problem to that of Theorem 3.1 would be: Given a certain stealthiness level in output, what is the maximum distortion in state that can be achieved by the attacker? And what is the corresponding attack? The following corollary answers these questions.

Corollary 3.1 *Consider the dynamical system under injection attacks depicted in Fig. 3.2. Then, in order for the attacker to ensure that the KL divergence rate between the original output and the attacked output is upper bounded by a (positive) constant R as*

$$\mathrm{KL}_\infty\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right) \leq R, \tag{3.13}$$

the maximum state distortion $\mathbb{E}\left[(\widehat{\mathbf{x}}_k - \mathbf{x}_k)^2\right]$ that can be achieved is given by

$$\sup_{\mathrm{KL}_\infty\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right) \leq R} \mathbb{E}\left[(\widehat{\mathbf{x}}_k - \mathbf{x}_k)^2\right] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{c^2} \left[\frac{\zeta S_{\mathbf{y}}^2(\omega)}{1 - \zeta S_{\mathbf{y}}(\omega)}\right] d\omega, \tag{3.14}$$

where ζ is the unique constant that satisfies

$$\frac{1}{2\pi}\int_{0}^{2\pi}\frac{1}{2}\left\{\frac{\frac{\zeta S_{\mathbf{y}}^{2}(\omega)}{1-\zeta S_{\mathbf{y}}(\omega)}}{S_{\mathbf{y}}(\omega)}-\ln\left[1+\frac{\frac{\zeta S_{\mathbf{y}}^{2}(\omega)}{1-\zeta S_{\mathbf{y}}(\omega)}}{S_{\mathbf{y}}(\omega)}\right]\right\}\mathrm{d}\omega$$

$$=\frac{1}{2\pi}\int_{0}^{2\pi}\frac{1}{2}\left\{\frac{\zeta S_{\mathbf{y}}(\omega)}{1-\zeta S_{\mathbf{y}}(\omega)}-\ln\left[\frac{1}{1-\zeta S_{\mathbf{y}}(\omega)}\right]\right\}\mathrm{d}\omega=R, \tag{3.15}$$

while

$$0<\zeta<\min_{\omega}\frac{1}{S_{\mathbf{y}}(\omega)}. \tag{3.16}$$

Note that herein $S_{\mathbf{y}}(\omega)$ is given by (3.8). Moreover, this maximum distortion is achieved when the attack signal $\{\mathbf{n}_k\}$ is chosen as a stationary colored Gaussian process with power spectrum

$$S_{\mathbf{n}}(\omega)=\frac{\left|\mathrm{e}^{\mathrm{j}\omega}-a\right|^{2}}{b^{2}c^{2}}\frac{\zeta S_{\mathbf{y}}^{2}(\omega)}{1-\zeta S_{\mathbf{y}}(\omega)}. \tag{3.17}$$

3.3.2 Feedback Control Systems

We will now proceed to examine (closed-loop) feedback control systems in this subsection. Specifically, consider the feedback control system depicted in Fig. 3.4, where the state-space model of the plant is given by

$$\begin{cases}\mathbf{x}_{k+1}=a\mathbf{x}_k+b\mathbf{u}_k+\mathbf{w}_k,\\ \mathbf{y}_k=c\mathbf{x}_k+\mathbf{v}_k,\end{cases}$$

while $K(z)$ is the transfer function of the (dynamic) output controller. Herein, $\mathbf{x}_k\in\mathbb{R}$ is the plant state, $\mathbf{u}_k\in\mathbb{R}$ is the plant input, $\mathbf{y}_k\in\mathbb{R}$ is the plant output, $\mathbf{w}_k\in\mathbb{R}$ is the

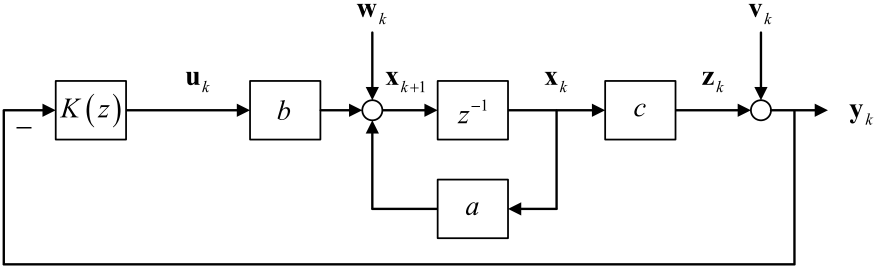

Fig. 3.4 A feedback control system

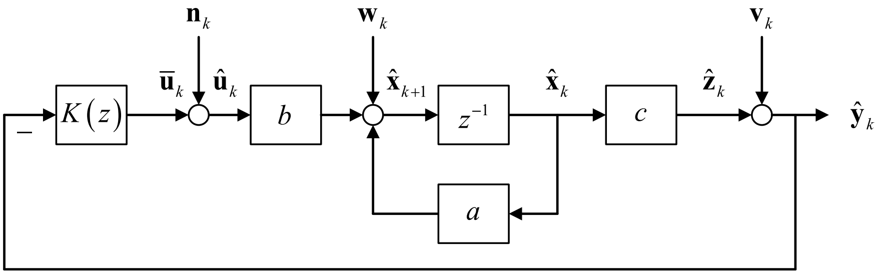

Fig. 3.5 A feedback control system under actuator attack

process noise, and $\mathbf{v}_k \in \mathbb{R}$ is the measurement noise. The system parameters are $a \in \mathbb{R}$, $b \in \mathbb{R}$, and $c \in \mathbb{R}$. Note that the plant is not necessarily stable. Meanwhile, we assume that $b, c \neq 0$, i.e., the plant is controllable and observable, and thus can be stabilized by controller $K(z)$. On the other hand, the transfer function of the plant is given by

$$P(z) = \frac{bc}{z-a}. \tag{3.18}$$

Suppose that $\{\mathbf{w}_k\}$ and $\{\mathbf{v}_k\}$ are stationary white Gaussian with variances $\sigma_{\mathbf{w}}^2$ and $\sigma_{\mathbf{v}}^2$, respectively. Furthermore, $\{\mathbf{w}_k\}$, $\{\mathbf{v}_k\}$, and $\mathbf{x}_0$ are assumed to be mutually independent. Assume also that $K(z)$ stabilizes $P(z)$, i.e., the closed-loop system is stable. Accordingly, $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$ are both stationary, and denote their power spectra by $S_{\mathbf{x}}(\omega)$ and $S_{\mathbf{y}}(\omega)$, respectively.

Consider then the scenario that an attack signal $\{\mathbf{n}_k\}$, $\mathbf{n}_k \in \mathbb{R}$, is to be added to the input of the plant $\{\mathbf{u}_k\}$ to deviate the plant state, while aiming to be stealthy in the plant output; see the depiction in Fig. 3.5. In fact, this corresponds to actuator attack. Note in particular that since we are now considering a closed-loop system, the presence of $\{\mathbf{n}_k\}$ will eventually distort the original $\{\mathbf{u}_k\}$ (through feedback) as well, which is an essential difference form the open-loop system setting considered in Sect. 3.3.1, and the distorted $\{\mathbf{u}_k\}$ will be denoted as $\{\overline{\mathbf{u}}_k\}$. In addition, we denote the true plant input under attack as $\{\widehat{\mathbf{u}}_k\}$, where

$$\widehat{\mathbf{u}}_k = \overline{\mathbf{u}}_k + \mathbf{n}_k, \tag{3.19}$$

whereas the plant under attack $\{\mathbf{n}_k\}$ is given by

$$\begin{cases} \widehat{\mathbf{x}}_{k+1} = a\widehat{\mathbf{x}}_k + b\widehat{\mathbf{u}}_k + \mathbf{w}_k = a\widehat{\mathbf{x}}_k + b\overline{\mathbf{u}}_k + b\mathbf{n}_k + \mathbf{w}_k, \\ \widehat{\mathbf{y}}_k = c\widehat{\mathbf{x}}_k + \mathbf{v}_k. \end{cases} \tag{3.20}$$

Meanwhile, suppose that the attack signal $\{\mathbf{n}_k\}$ is independent of $\{\mathbf{w}_k\}$, $\{\mathbf{v}_k\}$, and $\mathbf{x}_0$; consequently, $\{\mathbf{n}_k\}$ is independent of $\{\mathbf{x}_k\}$ and $\{\mathbf{y}_k\}$ as well.

The following theorem, as the second main result of this chapter, characterizes the fundamental trade-off between the distortion in state and the stealthiness in output for feedback control systems.

Theorem 3.2 *Consider the feedback control system under injection attacks depicted in Fig. 3.5. Suppose that the attacker needs to design the attack signal* $\{\mathbf{n}_k\}$ *to satisfy the following attack goal in terms of state distortion:*

$$\mathbb{E}\left[(\widehat{\mathbf{x}}_k - \mathbf{x}_k)^2\right] \geq D. \tag{3.21}$$

Then, the minimum KL divergence rate between the original output and the attacked output is given by

$$\inf_{\mathbb{E}\left[(\widehat{\mathbf{x}}_k - \mathbf{x}_k)^2\right] \geq D} \mathrm{KL}_\infty \left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right) = \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{2} \left\{ \frac{S_{\widehat{\mathbf{n}}}(\omega)}{S_{\mathbf{y}}(\omega)} - \ln\left[1 + \frac{S_{\widehat{\mathbf{n}}}(\omega)}{S_{\mathbf{y}}(\omega)}\right]\right\} \mathrm{d}\omega, \tag{3.22}$$

where

$$S_{\widehat{\mathbf{n}}}(\omega) = \frac{\zeta S_{\mathbf{y}}^2(\omega)}{1 - \zeta S_{\mathbf{y}}(\omega)}, \tag{3.23}$$

and $S_{\mathbf{y}}(\omega)$ *is given by*

$$S_{\mathbf{y}}(\omega) = \left| \frac{c}{\mathrm{e}^{\mathrm{j}\omega} - a + K\left(\mathrm{e}^{\mathrm{j}\omega}\right) bc} \right|^2 \sigma_{\mathbf{w}}^2 + \left| \frac{\mathrm{e}^{\mathrm{j}\omega} - a}{\mathrm{e}^{\mathrm{j}\omega} - a + K\left(\mathrm{e}^{\mathrm{j}\omega}\right) bc} \right|^2 \sigma_{\mathbf{v}}^2. \tag{3.24}$$

Herein, ζ *is the unique constant that satisfies*

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\zeta S_{\mathbf{y}}^2(\omega)}{1 - \zeta S_{\mathbf{y}}(\omega)} \mathrm{d}\omega = c^2 D, \tag{3.25}$$

while

$$0 < \zeta < \min_{\omega} \frac{1}{S_{\mathbf{y}}(\omega)}. \tag{3.26}$$

Moreover, the worst-case attack $\{\mathbf{n}_k\}$ *is a stationary colored Gaussian process with power spectrum*

$$S_{\mathbf{n}}(\omega) = \left| \frac{\mathrm{e}^{\mathrm{j}\omega} - a + K\left(\mathrm{e}^{\mathrm{j}\omega}\right) bc}{bc} \right|^2 \frac{\zeta S_{\mathbf{y}}^2(\omega)}{1 - \zeta S_{\mathbf{y}}(\omega)}. \tag{3.27}$$

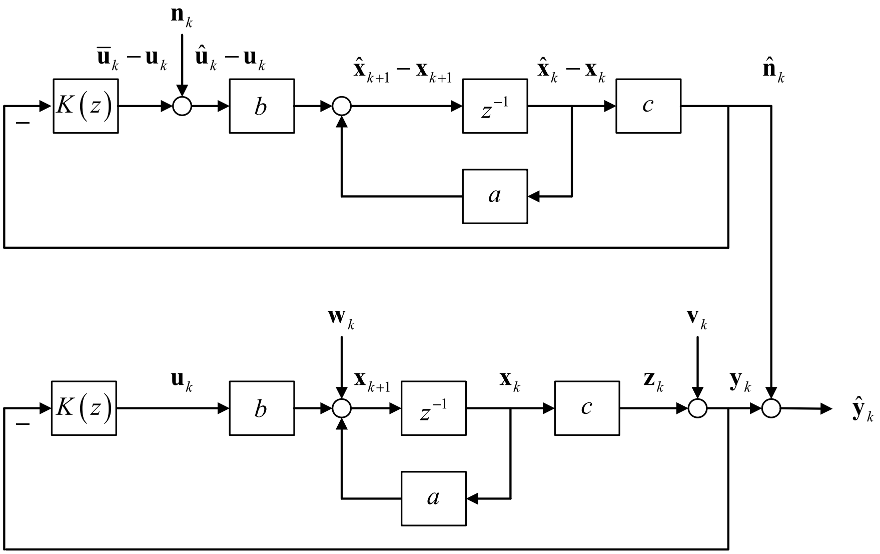

Fig. 3.6 A feedback control system under actuator attack: equivalent system

Proof Note first that when the closed-loop system is stable, the power spectrum of $\{\mathbf{y}_k\}$ is given by

$$\begin{aligned} S_{\mathbf{y}}(\omega) &= \frac{1}{b^2}\left|\frac{P\left(\mathrm{e}^{\mathrm{j}\omega}\right)}{1+K\left(\mathrm{e}^{\mathrm{j}\omega}\right)P\left(\mathrm{e}^{\mathrm{j}\omega}\right)}\right|^2 \sigma_{\mathbf{w}}^2 + \left|\frac{1}{1+K\left(\mathrm{e}^{\mathrm{j}\omega}\right)P\left(\mathrm{e}^{\mathrm{j}\omega}\right)}\right|^2 \sigma_{\mathbf{v}}^2, \\ &= \frac{1}{b^2}\left|\frac{\frac{bc}{\mathrm{e}^{\mathrm{j}\omega}-a}}{1+K\left(\mathrm{e}^{\mathrm{j}\omega}\right)\frac{bc}{\mathrm{e}^{\mathrm{j}\omega}-a}}\right|^2 \sigma_{\mathbf{w}}^2 + \left|\frac{1}{1+K\left(\mathrm{e}^{\mathrm{j}\omega}\right)\frac{bc}{\mathrm{e}^{\mathrm{j}\omega}-a}}\right|^2 \sigma_{\mathbf{v}}^2, \\ &= \left|\frac{c}{\mathrm{e}^{\mathrm{j}\omega}-a+K\left(\mathrm{e}^{\mathrm{j}\omega}\right)bc}\right|^2 \sigma_{\mathbf{w}}^2 + \left|\frac{\mathrm{e}^{\mathrm{j}\omega}-a}{\mathrm{e}^{\mathrm{j}\omega}-a+K\left(\mathrm{e}^{\mathrm{j}\omega}\right)bc}\right|^2 \sigma_{\mathbf{v}}^2. \end{aligned}$$

Note then that since the systems are linear, the system in Fig. 3.5 is equivalent to that of Fig. 3.6, where

$$\widehat{\mathbf{y}}_k = \mathbf{y}_k + \widehat{\mathbf{n}}_k,$$

and $\{\widehat{\mathbf{n}}_k\}$ is the output of the closed-loop system composed by the controller $K(z)$ and the plant

$$\begin{cases} \widehat{\mathbf{x}}_{k+1} - \mathbf{x}_{k+1} = a\left(\widehat{\mathbf{x}}_k - \mathbf{x}_k\right) + b\left(\overline{\mathbf{u}}_k - \mathbf{u}_k\right) + b\mathbf{n}_k, \\ \widehat{\mathbf{n}}_k = c\left(\widehat{\mathbf{x}}_k - \mathbf{x}_k\right), \end{cases}$$

as depicted by the upper half of Fig. 3.6. Meanwhile, as in the case of Fig. 3.3, the system in Fig. 3.6 may also be viewed as a “virtual channel” modeled as

$$\widehat{\mathbf{y}}_k = \mathbf{y}_k + \widehat{\mathbf{n}}_k$$

with noise constraint

$$\mathbb{E}\left[\widehat{\mathbf{n}}_k^2\right] \geq c^2 D,$$

where $\{\mathbf{y}_k\}$ is the channel input, $\{\widehat{\mathbf{y}}_k\}$ is the channel output, and $\{\widehat{\mathbf{n}}_k\}$ is the channel noise that is independent of $\{\mathbf{y}_k\}$. Then, following procedures similar to those in the proof of Theorem 3.1, it can be derived that

$$\inf_{\mathbb{E}\left[(\widehat{\mathbf{x}}_k - \mathbf{x}_k)^2\right] \geq D} \mathrm{KL}_{\infty}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right) = \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{2} \left\{ \frac{S_{\widehat{\mathbf{n}}}(\omega)}{S_{\mathbf{y}}(\omega)} - \ln\left[1 + \frac{S_{\widehat{\mathbf{n}}}(\omega)}{S_{\mathbf{y}}(\omega)}\right] \right\} \mathrm{d}\omega,$$

where

$$S_{\widehat{\mathbf{n}}}(\omega) = \frac{\zeta S_{\mathbf{y}}^2(\omega)}{1 - \zeta S_{\mathbf{y}}(\omega)},$$

and ζ is the unique constant that satisfies

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} S_{\widehat{\mathbf{n}}}(\omega)\, \mathrm{d}\omega = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\zeta S_{\mathbf{y}}^2(\omega)}{1 - \zeta S_{\mathbf{y}}(\omega)} \mathrm{d}\omega = c^2 D,$$

while

$$0 < \zeta < \min_{\omega} \frac{1}{S_{\mathbf{y}}(\omega)}.$$

In addition, since

$$\begin{aligned} S_{\widehat{\mathbf{n}}}(\omega) &= \left| \frac{P\left(\mathrm{e}^{\mathrm{j}\omega}\right)}{1 + K\left(\mathrm{e}^{\mathrm{j}\omega}\right) P\left(\mathrm{e}^{\mathrm{j}\omega}\right)} \right|^2 S_{\mathbf{n}}(\omega) = \left| \frac{\frac{bc}{\mathrm{e}^{\mathrm{j}\omega} - a}}{1 + K\left(\mathrm{e}^{\mathrm{j}\omega}\right) \frac{bc}{\mathrm{e}^{\mathrm{j}\omega} - a}} \right|^2 S_{\mathbf{n}}(\omega), \\ &= \left| \frac{bc}{\mathrm{e}^{\mathrm{j}\omega} - a + K\left(\mathrm{e}^{\mathrm{j}\omega}\right) bc} \right|^2 S_{\mathbf{n}}(\omega), \end{aligned}$$

we have

$$S_{\mathbf{n}}(\omega) = \left| \frac{\mathrm{e}^{\mathrm{j}\omega} - a + K\left(\mathrm{e}^{\mathrm{j}\omega}\right) bc}{bc} \right|^2 S_{\widehat{\mathbf{n}}}(\omega) = \left| \frac{\mathrm{e}^{\mathrm{j}\omega} - a + K\left(\mathrm{e}^{\mathrm{j}\omega}\right) bc}{bc} \right|^2 \frac{\zeta S_{\mathbf{y}}^2(\omega)}{1 - \zeta S_{\mathbf{y}}(\omega)}.$$

This concludes the proof. ■

It is worth mentioning that the $S_{\mathbf{y}}(\omega)$ for Theorem 3.2 is given by (3.24), which differs significantly from that given by (3.8) for Theorem 3.1, although the notations are the same. Accordingly, η, $S_{\mathbf{n}}(\omega)$, and so on, will all be different between the two cases in spite of the same notations.

Note also that $S_{\mathbf{n}}(\omega)$ can be rewritten as

$$S_{\mathbf{n}}(\omega) = \left| \frac{1 + K\left(\mathrm{e}^{\mathrm{j}\omega}\right) P\left(\mathrm{e}^{\mathrm{j}\omega}\right)}{P\left(\mathrm{e}^{\mathrm{j}\omega}\right)} \right|^2 \frac{\zeta S_{\mathbf{y}}^2(\omega)}{1 - \zeta S_{\mathbf{y}}(\omega)}, \tag{3.28}$$

which indicates that the attacker only needs to know the power spectrum of the original system output $\{\mathbf{y}_k\}$ and the transfer function of the closed-loop system (from $\{\mathbf{n}_k\}$ to $\{\widehat{\mathbf{y}}_k\}$), i.e.,

$$\frac{P(z)}{1 + K(z) P(z)}, \tag{3.29}$$

in order to carry out this worst-case attack.

Again, we may examine the dual problem as follows.

Corollary 3.2 *Consider the feedback control system under injection attacks depicted in Fig. 3.5. Then, in order for the attacker to ensure that the KL divergence rate between the original output and the attacked output is upper bounded by a (positive) constant R as*

$$\mathrm{KL}_{\infty}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right) \leq R, \tag{3.30}$$

the maximum state distortion $\mathbb{E}\left[(\widehat{\mathbf{x}}_k - \mathbf{x}_k)^2\right]$ that can be achieved is given by

$$\sup_{\mathrm{KL}_{\infty}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right) \leq R} \mathbb{E}\left[(\widehat{\mathbf{x}}_k - \mathbf{x}_k)^2\right] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{c^2} \left[\frac{\zeta S_{\mathbf{y}}^2(\omega)}{1 - \zeta S_{\mathbf{y}}(\omega)} \right] \mathrm{d}\omega, \tag{3.31}$$

where ζ satisfies

$$\begin{aligned} &\frac{1}{2\pi} \int_{0}^{2\pi} \frac{1}{2} \left\{ \frac{\frac{\zeta S_{\mathbf{y}}^2(\omega)}{1-\zeta S_{\mathbf{y}}(\omega)}}{S_{\mathbf{y}}(\omega)} - \ln \left[1 + \frac{\frac{\zeta S_{\mathbf{y}}^2(\omega)}{1-\zeta S_{\mathbf{y}}(\omega)}}{S_{\mathbf{y}}(\omega)} \right] \right\} \mathrm{d}\omega \\ &\quad = \frac{1}{2\pi} \int_{0}^{2\pi} \frac{1}{2} \left\{ \frac{\zeta S_{\mathbf{y}}(\omega)}{1 - \zeta S_{\mathbf{y}}(\omega)} - \ln \left[\frac{1}{1 - \zeta S_{\mathbf{y}}(\omega)} \right] \right\} \mathrm{d}\omega = R, \end{aligned} \tag{3.32}$$

while

$$0 < \zeta < \min_{\omega} \frac{1}{S_{\mathbf{y}}(\omega)}. \tag{3.33}$$

Note that herein $S_{\mathbf{y}}(\omega)$ is given by (3.24). Moreover, this maximum distortion is achieved when the attack signal $\{\mathbf{n}_k\}$ is chosen as a stationary colored Gaussian process with power spectrum

$$S_{\mathbf{n}}(\omega) = \left| \frac{\mathrm{e}^{\mathrm{j}\omega} - a + K\left(\mathrm{e}^{\mathrm{j}\omega}\right) bc}{bc} \right|^2 \frac{\zeta S_{\mathbf{y}}^2(\omega)}{1 - \zeta S_{\mathbf{y}}(\omega)}. \tag{3.34}$$

3.4 Simulation

In this section, we will utilize (toy) numerical examples to illustrate the fundamental stealthiness–distortion trade-offs in linear Gaussian open-loop dynamical systems as well as (closed-loop) feedback control systems.

Consider first open-loop dynamical systems as in Sect. 3.3.1. Let $a = 0.5, b = 1, c = 1, \sigma_{\mathbf{w}}^2 = 1, \sigma_{\mathbf{v}}^2 = 1$, and $S_{\mathbf{u}}(\omega) = 1$ therein for simplicity. Accordingly, we have

$$S_{\mathbf{y}}(\omega) = \frac{2}{\left|\mathrm{e}^{\mathrm{j}\omega} - 0.5\right|^2} + 1 = \frac{2}{(\cos\omega - 0.5)^2 + \sin^2\omega} + 1.$$

In such a case, the relation between the minimum KL divergence rate $\mathrm{KL}_{\infty}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right)$ (denoted as KL in the figure) and the distortion bound D is illustrated in Fig. 3.7. It is clear that KL increases (strictly) with D, i.e., in order for the attacker to achieve larger distortion, the stealthiness level of the attack will inevitably decrease.

Note that the relation between the maximum distortion $\mathbb{E}\left[(\widehat{\mathbf{x}}_k - \mathbf{x}_k)^2\right]$ and the KL divergence rate bound R in Corollary 3.1 is essentially the same as that between the distortion bound D and the minimum KL divergence rate $\mathrm{KL}_{\infty}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right)$ in Theorem 3.1.

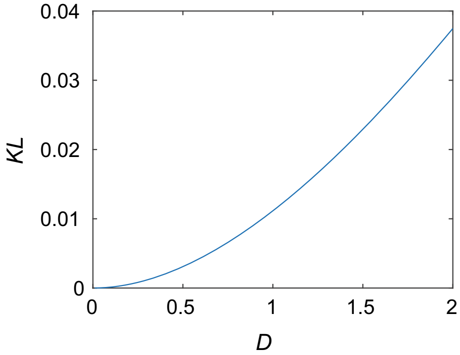

Fig. 3.7 The relation between $\mathrm{KL}_{\infty}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right)$ (denoted as KL) and D in Open-Loop Dynamical Systems

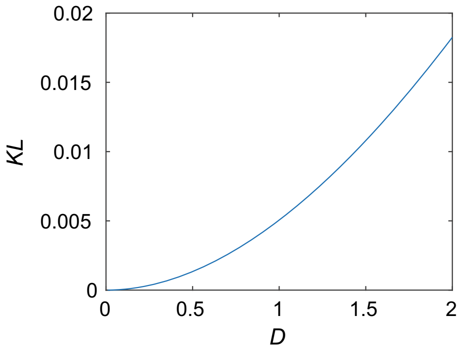

Fig. 3.8 The relation between $\mathrm{KL}_{\infty}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right)$ (denoted as KL) and D in Feedback Control Systems

Consider then feedback control systems as in Sect. 3.3.2. Let $a = 2$, $b = 1$, $c = 1$, $\sigma_{\mathbf{w}}^2 = 1$, $\sigma_{\mathbf{v}}^2 = 1$, and $K(z) = 2$ therein for simplicity. Accordingly, we have

$$S_{\mathbf{y}}(\omega) = 1 + \left|\mathrm{e}^{\mathrm{j}\omega} - 2\right|^2 = 1 + (\cos\omega - 2)^2 + \sin^2\omega.$$

In such a case, the relation between the minimum KL divergence rate $\mathrm{KL}_{\infty}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right)$ (denoted as KL in the figure) and the distortion bound D is illustrated in Fig. 3.8. Again, KL increases (strictly) with D, whereas the relationship between the maximum distortion $\mathbb{E}\left[(\widehat{\mathbf{x}}_k - \mathbf{x}_k)^2\right]$ and the KL divergence rate bound R in Corollary 3.2 is essentially the same as that between the distortion bound D and the minimum KL divergence rate $\mathrm{KL}_{\infty}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right)$ in Theorem 3.2.

3.5 Conclusion

In this chapter, we have presented the fundamental stealthiness–distortion trade-offs of linear Gaussian open-loop dynamical systems and (closed-loop) feedback control systems under data injection attacks, and explicit formulas have been obtained in terms of power spectra that characterize analytically the stealthiness–distortion trade-offs as well as the properties of the worst-case attacks.

So why do we care about explicit formulas in the first place? One value of the explicit stealthiness–distortion trade-off formula for feedback control systems, for instance, is that they render the subsequent controller design explicit (and intuitive) as well. To be more specific, given a threshold on the output stealthiness, it is already known from Corollary 3.2 what the maximum distortion in state that can be achieved by the attacker is. Then, one natural control design criterion will be to design the controller $K(z)$ so as to minimize this maximum distortion. Mathematically, this minimax problem can be formulated as follows:

$$\inf_{K(z)} \sup_{\mathrm{KL}_{\infty}\left(p_{\widehat{\mathbf{y}}} \| p_{\mathbf{y}}\right) \leq R} \mathbb{E}\left[\left(\widehat{\mathbf{x}}_k - \mathbf{x}_k\right)^2\right] = \inf_{K(z)} \left\{ \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{1}{c^2} \left[\frac{\zeta S_{\mathbf{y}}^2(\omega)}{1 - \zeta S_{\mathbf{y}}(\omega)} \right] \mathrm{d}\omega \right\},$$

where

$$\begin{aligned} S_{\mathbf{y}}(\omega) &= \left| \frac{c}{\mathrm{e}^{\mathrm{j}\omega} - a + K\left(\mathrm{e}^{\mathrm{j}\omega}\right) bc} \right|^2 \sigma_{\mathbf{w}}^2 + \left| \frac{\mathrm{e}^{\mathrm{j}\omega} - a}{\mathrm{e}^{\mathrm{j}\omega} - a + K\left(\mathrm{e}^{\mathrm{j}\omega}\right) bc} \right|^2 \sigma_{\mathbf{v}}^2, \\ &= \left| \frac{P\left(\mathrm{e}^{\mathrm{j}\omega}\right)}{1 + K\left(\mathrm{e}^{\mathrm{j}\omega}\right) P\left(\mathrm{e}^{\mathrm{j}\omega}\right)} \right|^2 \frac{\sigma_{\mathbf{w}}^2}{b^2} + \left| \frac{1}{1 + K\left(\mathrm{e}^{\mathrm{j}\omega}\right) P\left(\mathrm{e}^{\mathrm{j}\omega}\right)} \right|^2 \sigma_{\mathbf{v}}^2, \end{aligned}$$

whereas the infimum is taken over all $K(z)$ that stabilizes the plant $P(z)$. Herein, ζ can be treated as a tuning parameter as long as it satisfies

$$0 < \zeta < \min_{\omega} \frac{1}{S_{\mathbf{y}}(\omega)}.$$

We will, however, leave more detailed investigations of this formulation to future research.

Other potential future research directions include the investigation of such trade-offs for state estimation systems. It might also be interesting to examine the security–privacy trade-offs (see, e.g., Farokhi and Esfahani (2018), Fang and Zhu (2020, 2021)).

References

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

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

P. Cheng, L. Shi, B. Sinopoli, Guest editorial special issue on secure control of cyber-physical systems. IEEE Trans. Control Netw. Syst. **4**(1), 1–3 (2017)

M.S. Chong, H. Sandberg, A.M. Teixeira, A tutorial introduction to security and privacy for cyber-physical systems, in *Proceedings of the European Control Conference (ECC)* (2019), pp. 968–978

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

S.M. Dibaji, M. Pirani, D.B. Flamholz, A.M. Annaswamy, K.H. Johansson, A. Chakrabortty, A systems and control perspective of CPS security. Ann. Rev. Control **47**, 394–411 (2019)

S. Fang, J. Chen, H. Ishii, *Towards Integrating Control and Information Theories: From Information-Theoretic Measures to Control Performance Limitations* (Springer, 2017)

S. Fang, Q. Zhu, Channel leakage, information-theoretic limitations of obfuscation, and optimal privacy mask design for streaming data (2020), arXiv:2008.04893

S. Fang, Q. Zhu, Fundamental limits of obfuscation for linear Gaussian dynamical systems: an information-theoretic approach, in *Proceedings of the American Control Conference* (2021)

S. Fang, Q. Zhu, Fundamental stealthiness-distortion tradeoffs in dynamical systems under injection attacks: a power spectral analysis, in *Proceedings of the European Control Conference* (2021)

S. Fang, Q. Zhu, Independent Gaussian distributions minimize the Kullback–Leibler (KL) divergence from independent Gaussian distributions (2020), arXiv: 2011.02560

F. Farokhi, P.M. Esfahani, Security versus privacy, in *Proceedings of the IEEE Conference on Decision and Control* (2018), pp. 7101–7106

J. Giraldo, D. Urbina, A. Cardenas, J. Valente, M. Faisal, J. Ruths, N.O. Tippenhauer, H. Sandberg, R. Candell, A survey of physics-based attack detection in cyber-physical systems. ACM Comput. Surv. (CSUR) **51**(4), 76 (2018)

I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, *Deep Learning* (MIT Press, 2016)

U. Grenander, G. Szegö, *Toeplitz Forms and Their Applications* (University of California Press, 1958)

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

J. Gutiérrez-Gutiérrez, P.M. Crespo, Asymptotically equivalent sequences of matrices and Hermitian block Toeplitz matrices with continuous symbols: applications to MIMO systems. IEEE Trans. Inf. Theory **54**(12), 5671–5680 (2008)

K.H. Johansson, G.J. Pappas, P. Tabuada, C.J. Tomlin, Guest editorial special issue on control of cyber-physical systems. IEEE Trans. Autom. Control **59**(12), 3120–3121 (2014)

S.M. Kay, *Information-Theoretic Signal Processing and its Applications* (Sachuest Point Publishers, 2020)

S. Kullback, *Information Theory and Statistics* (Courier Corporation, 1997)

S. Kullback, R.A. Leibler, On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)

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

A. Lindquist, G. Picci, *Linear Stochastic Systems: A Geometric Approach to Modeling*. Estimation and Identification. (Springer, 2015)

L. Ljung, *System Identification: Theory For the User* (Prentice Hall, 1999)

L. Pardo, *Statistical Inference Based on Divergence Measures* (CRC Press, 2006)

M.S. Pinsker, *Information and Information Stability of Random Variables and Processes* (Holden Day, San Francisco, CA, 1964)

H.V. Poor, *An Introduction to Signal Detection and Estimation* (Springer, 2013)

R. Poovendran, K. Sampigethaya, S.K.S. Gupta, I. Lee, K.V. Prasad, D. Corman, J.L. Paunicka, Special issue on cyber-physical systems [scanning the issue]. Proc. IEEE **100**(1), 6–12 (2012)

H. Sandberg, S. Amin, K.H. Johansson, Cyberphysical security in networked control systems: an introduction to the issue. IEEE Control Syst. Mag. **35**(1), 20–23 (2015)

P. Stoica, R. Moses, *Spectral Analysis of Signals* (Prentice Hall, 2005)

A. Stoorvogel, J. Van Schuppen, System identification with information theoretic criteria, in *Identification, Adaptation, Learning: The Science of Learning Models from Data*, ed. by S. Bittanti, G. Picci (Springer, 1996)

S. Weerakkody, O. Ozel, Y. Mo, B. Sinopoli, Resilient control in cyber-physical systems: countering uncertainty, constraints, and adversarial behavior, Foundations and Trends®. Syst. Control **7**(1–2), 1–252 (2019)

R. Zhang, P. Venkitasubramaniam, Stealthy control signal attacks in linear quadratic Gaussian control systems: detectability reward tradeoff. IEEE Trans. Inf. Foren. Secur. **12**(7), 1555–1570 (2017)