Security metrics for control systems

André M.H. Teixeira

Abstract In this chapter, we consider stealthy cyber- and physical attacks against control systems, where malicious adversaries aim at maximizing the impact on control performance, while simultaneously remaining undetected. As an initial goal, we develop security-related metrics to quantify the impact of stealthy attacks on the system. The key novelty of these metrics is that they jointly consider impact and detectability of attacks, unlike classical sensitivity metrics in robust control and fault detection. The final objective of this work is to use such metrics to guide the design of optimal resilient controllers and detectors against stealthy attacks, akin to the classical design of optimal robust controllers. We report preliminary investigations on the design of resilient observer-based controllers and detectors, which are supported and illustrated through numerical examples.

1 Introduction

Cyber-security of control systems has been receiving increasing attention in recent years. Overviews of existing cyber-threats and vulnerabilities in networked control systems has been presented by different authors (Cárdenas et al, 2008; Teixeira et al, 2015b). Adversaries endowed with rationality and intent are highlighted as one of the key items in security for control systems, as opposed to natural faults and disturbances. Therefore, these adversaries may exploit existing vulnerabilities and limitations in anomaly detection mechanisms and remain undetected. In fact, Pasqualetti et al (2013) uses tools from geometric control to study such fundamental limitations and characterizes a set of stealthy attack policies for networked systems modeled by differential-algebraic equations. Related stealthy attack policies were also considered in Smith (2015); Teixeira et al (2015b), while the work by Fawzi et al (2014)

André M.H. Teixeira
Department of Electrical Engineering, Uppsala University, Sweden e-mail: andre.teixeira@angstrom.uu.se

characterizes the number of corrupted sensor channels that cannot be detected during a finite time-interval. A common thread within these approaches is that stealthy attacks are constrained to be entirely decoupled from the anomaly detector's output. Classes of attacks that are in theory detectable, but hard to detect in practice, have not received as much attention.

Another important direction is to analyze the potential damage of stealthy attacks. Recently, Bai et al (2017) investigated the detectability limitations and performance degradation of data injection attacks in stochastic control systems. The impact of stealthy data injection attacks on sensors is also investigated in Mo and Sinopoli (2016), which characterized the set of states reachable by stealthy adversary. The work in Teixeira et al (2015c) formulated the impact of data injection attacks in finite time-horizon as a generalized eigenvalue problem, whereas Umsonst et al (2017) considered an alternative formulation that allowed for the impact to be characterized as the solution to a convex optimization problem. A similar approach was considered in Shames et al (2017) for impulsive attacks.

While this set of results is useful to assess the impact of stealthy cyber-attacks on control systems, they cannot be used to directly design more resilient controllers, since the optimization problems have a complex non-convex dependence on the design parameters.

The impact and detectability of data injection attacks on discrete-time has also been jointly considered in the author's previous work (Teixeira et al, 2015a). The impact of stealthy attacks is characterized as the solution to a convex problem that has a remarkable similarity with existing optimization-based techniques to design optimal $\mathscr{H}_\infty$ robust detectors and controllers (Wang et al, 2007; Scherer and Weiland, 2010).

As main contributions of this chapter, we revisit the control system security problem in Teixeira et al (2015a), but now for continuous-time systems. The goal is to investigate possible metrics with which to analyze the security of control systems to malicious adversaries that aim at maximizing impact while simultaneously minimizing detection. Classical metrics in robust control and fault detection are revisited, namely the $\mathscr{H}_\infty$ norm Zhou et al (1996) and the $\mathscr{H}_-$ index (Wang et al, 2007). Their suitability to security analysis is discussed, from which we conclude that they have limited applicability to security, as they consider impact and detection separately. Then, a first heuristic approach to integrate these metrics together is taken and further examined, which also shares some of the limitations of the classical metrics.

The continuous-time version of the security metric developed in Teixeira et al (2015a) is then presented: the output-to-output gain. Results characterizing the metric are given, based on which an efficient computation approach is proposed. Furthermore, fundamental limitations of this metric are also characterized, which are aligned with well-known fundamental limitations in the detectability of attacks (Pasqualetti et al, 2013; Bai et al, 2017).

Finally, a first attempt to use this security metric in the design of optimal resilient controllers and detectors against stealthy attacks are investigated. An heuristic method to search for sub-optimal solutions is presented, based on alternating minimization.

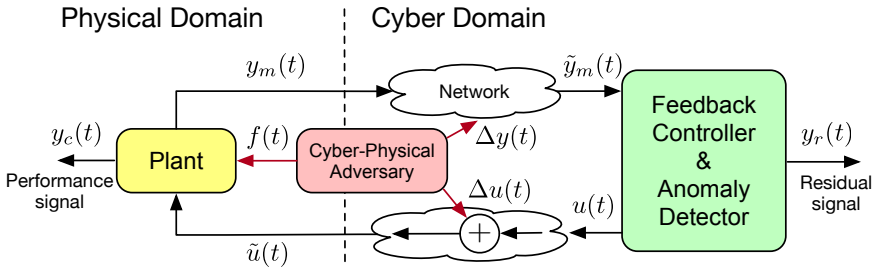

Fig. 1 Networked control system under cyber-physical attacks.

The discussion and results in this chapter are illustrated and supported by numerical experiments on a single closed-loop system, with small examples distributed through the different sections of the chapter.

2 Closed-loop system under cyber and physical attacks

This section details the attack scenario and the closed-loop system structure under study, following a similar modeling framework as in Teixeira et al (2015b).

Consider a control system depicted with a physical plant (**P**) controlled over a communication network, and an integrated observer-based feedback controller and anomaly detector (**F**). The closed-loop system under attacks is depicted in Figure 1, and its dynamics are described by the following equations:

$$\begin{aligned}
\mathbf{P}: &\begin{cases} \dot{x}_p(t) = A_p x_p(t) + B_p \tilde{u}(t) + E_p f(t) \\ y_c(t) = C_p x_p(t) + D_p \tilde{u}(t) \\ y_m(t) = C_m x(t) \end{cases} \\
\mathbf{F}: &\begin{cases} \dot{\hat{x}}_p(t) = A_p \hat{x}(t) + B_p u(t) + K y_r(t) \\ \hat{y}_m(t) = C_m \hat{x}(t) \\ u(t) = L\hat{x}(t) \\ y_r(t) = \tilde{y}_m(t) - \hat{y}_m(t), \end{cases}
\end{aligned} \tag{1}$$

where $x_p \in \mathbb{R}^{n_x}$ denotes the state of the plant, $y_m(t) \in \mathbb{R}^{n_y}$ is the measurement signal transmitted by the sensors from the plant to the observer, $u \in \mathbb{R}^{n_u}$ is the control signal computed by the observer-based controller and then transmitted to the actuator, and $f(t) \in \mathbb{R}^{n_f}$ is a physical fault signal, possibly inserted by a malicious adversary.

The level of performance of the closed-loop system over a given time-horizon $[0, T]$ is measured by the quadratic control cost

$$J(x(t),\tilde{u}(t))_{[0,T]} \triangleq \int_0^T \|y_c(t)\|_2^2\, dt \triangleq \|y_c\|^2_{\mathscr{L}_2[0,T]}. \tag{2}$$

The signal $y_c(t) \in \mathbb{R}^{n_c}$ is thus denoted as the performance output.

The communication network may be subject to malicious cyber-attacks, which are able to hijack, read, and re-write the data packets flowing through the network. The possibly compromised measurement and actuator signals at the corresponding receiver are denoted by $\tilde{y}_m(t)$ and $\tilde{u}(t)$, respectively.

The observer produces an estimate of the plant's state, $\hat{x}(t)$, which has a dual role. On the one hand, the estimate is the basis for computing the control input $u(t)$ that steers the system. On the other hand, the observer's estimate is also used to generate a so-called residual $y_r(t)$, which is evaluated to detect the presence of anomalies. Thus, the residual is also called as the detection output. In particular, we suppose that an anomaly is detected if the energy of the residual signal over a given time-horizon $[0, T]$ exceeds a certain threshold, *i.e.*,

$$\|y_r\|^2_{\mathscr{L}_2[0,T]} > \tau_r^2. \tag{3}$$

In the remainder of the chapter, without loss of generality, we let $\tau_r = 1$.

2.1 Attack scenario and adversary model

In this chapter, we consider the class of false-data injection attacks on data exchanged with sensors and actuators, possibly combined with a physical attack $f(t)$. The cyber-attack component can be modeled as additive corruptions of the sensor and actuator signals, described by

$$\begin{aligned} \tilde{u}(t) &= u(t) + \Delta u(t) & \Delta u(t) &= \Gamma_u a_u(t) \\ \tilde{y}_m(t) &= y_m(t) + \Delta y_m(t), & \Delta y_m(t) &= \Gamma_y a_y(t), \end{aligned} \tag{4}$$

where $a_u(t) \in \mathbb{R}^{m_u}$ and $a_y(t) \in \mathbb{R}^{m_y}$ are the data corruptions added to the actuator and sensor signals, respectively, and $\Gamma_u \in \mathbb{B}^{n_u \times m_u}$ and $\Gamma_y \in \mathbb{B}^{n_y \times m_y}$ are binary-valued matrices indicating which m_u and m_y channels can be corrupted by the adversary. Additionally, we consider that the adversary can also stage a physical attack through the fault signal $f(t)$, possibly in coordination with the cyber-attack on sensor and actuator data.

In terms of the adversary model, we consider the worst-case scenario where the adversary has perfect knowledge of the closed-loop system model in (1).

Moreover, concerning the intent of the malicious adversary, we suppose that the attacker has two objectives. First, the adversary aims at corrupting the sensor and actuator data so that the closed-loop system performance is deteriorated as much as possible. This means that the adversary aims at maximizing the control cost (2). Second, the attacker wishes to minimize the detection alarms (3), and therefore avoid being detected.

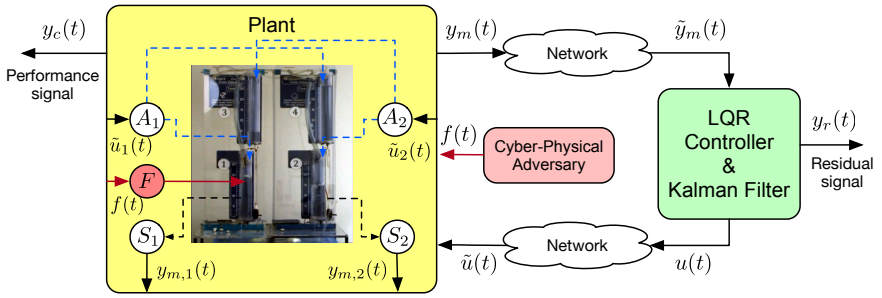

Fig. 2 The quadruple tank process, controlled through an LQG controller, under a physical attack. The physical attack may be modeled with $\Gamma_u = \Gamma_y = \emptyset$ and $E_p = e_1$.

Example 1. Throughout the chapter, the quadruple water tank system (Johansson et al, 1999) will be used as a reoccurring example. The plant is depicted in Figure 2, which illustrates one of the attack scenarios considered in this chapter.

The plant consists of four water tanks, with four states associated with the water level of each tank, two measurements signals corresponding to the water level in two of the tanks, and two actuators corresponding to two water pumps. Each water pump delivers water to two tanks, as depicted in Figure 2, The flow ratio from each pump to the respective tanks is determined by a valve. The valves are configured so that the plant possesses one unstable transmission zero from $\tilde{u}(t)$ to $y_m(t)$.

The observer-based controller is implemented as a LQG controller, by means of a Kalman Filter that feeds its state estimate to an LQR controller. The LQG controller is designed as to minimize the quadratic cost (2). The closed-loop system dynamics are described by (1), with the data:

$$A_p = \begin{bmatrix} -0.1068 & 0 & 0.0275 & 0 \\ 0 & -0.0903 & 0 & 0.0258 \\ 0 & 0 & -0.0275 & 0 \\ 0 & 0 & 0 & -0.0258 \end{bmatrix}, \quad B_p = \begin{bmatrix} 0.0802 & 0 \\ 0 & 0.0807 \\ 0 & 0.1345 \\ 0.1337 & 0 \end{bmatrix},$$

$$C_m = \begin{bmatrix} 0.2000 & 0 & 0 & 0 \\ 0 & 0.2000 & 0 & 0 \end{bmatrix},$$

$$C_p = \begin{bmatrix} 6.3246 & 0 & 0 & 0 \\ 0 & 6.3246 & 0 & 0 \\ 0 & 0 & 4.4721 & 0 \\ 0 & 0 & 0 & 4.4721 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad D_p = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 3.1623 & 0 \\ 0 & 3.1623 \end{bmatrix},$$

$$L = \begin{bmatrix} -0.5997 & -0.2224 & 0.1003 & -1.2153 \\ -0.1916 & -0.6606 & -1.1847 & 0.0816 \end{bmatrix}, \quad K = \begin{bmatrix} 0.0345 & 0.0150 \\ 0.0150 & 0.0414 \\ 0.0456 & 0.0633 \\ 0.0561 & 0.0515 \end{bmatrix}.$$

The closed-loop system is under attack on a subset of the sensors and actuators, possibly complemented with a physical attack that can affect each water tank independently. In particular, throughout the chapter we shall discuss different attack scenarios where the adversary corrupts at most two communication channels and physically attacks at most two water tanks. The former is modeled through Γ_u and Γ_y, while the latter is captured by having $E_p = [e_i \; e_k] \in \mathbb{R}^{4\times 2}$, where e_i is the i-th column of the identity matrix.

As a first example, suppose that the adversary corrupts both pumps, which is modeled by (4) with $\Gamma_u = I$ and $\Gamma_y = \emptyset$. In this case, it has been previously shown in the literature (Teixeira et al, 2012, 2015b) that an attack on the actuators mimicking the unstable transmission zero dynamics will have an arbitrarily large impact on the plant's states, while remaining undetectable. For our specific system, the attack signal could be designed as $a(t) = \delta e^{0.0178t}\left[-0.2243 \quad 0.2200\right]^\top$, for a sufficiently small scalar δ. Such an attack will have a negligible effect on the detection output, while driving the states, and thus the control cost, towards infinity.

2.2 Toward metrics for security analysis

Given the control system and attack scenario previously described, the next sections look into possible metrics for characterizing the worst-case attack, its impact on performance, and its level of detectability.

For simplicity, we will re-write the dynamics of the closed-loop system under attack in a more compact form. To this end, let us define the estimation error as $e(t) \triangleq x_p(t) - \hat{x}_p(t)$, the augmented state of the plant and observer as $x(t) \triangleq \left[x_p(t)^\top \; e(t)^\top\right]^\top$ and the augmented attack signal as $a(t) \triangleq \left[a_u(t)^\top \; a_y(t)^\top \; f(t)^\top\right]^\top \in \mathbb{R}^{n_a}$. The closed-loop dynamics (1) under the considered attack (4) are compactly described by

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Ba(t) \\ y_c(t) &= C_c x(t) + D_c a(t) \\ y_r(t) &= C_r x(t) + D_r a(t), \end{aligned} \tag{5}$$

where the matrices are given by

$$\begin{aligned} A &= \begin{bmatrix} A_p + B_p L & -B_p L \\ 0 & A_p + KC_m \end{bmatrix}, \quad B = \begin{bmatrix} B_p\Gamma_u & 0 & E_p \\ B_p\Gamma_u & K\Gamma_y & E_p \end{bmatrix}, \\ C_c &= \begin{bmatrix} C_p + D_p L & -D_p L \end{bmatrix}, \qquad D_c = \begin{bmatrix} D_p\Gamma_u & 0 & 0 \end{bmatrix}, \\ C_r &= \begin{bmatrix} 0 & C_m \end{bmatrix}, \qquad D_r = \begin{bmatrix} 0 & \Gamma_y & 0 \end{bmatrix}. \end{aligned} \tag{6}$$

Furthermore, we shall denote $\Sigma_c \triangleq (A, B, C_c, D_c)$ and $\Sigma_r \triangleq (A, B, C_r, D_r)$ as the realizations of the closed-loop system as seen from the perspectives of attack $a(t)$ to the outputs $y_c(t)$ and $y_r(t)$, respectively.

Note that the adversary model places no explicit constraint on the attack signal $a(t)$. Therefore, we will consider generic attack signals that lie in the so-called extended $\mathscr{L}_2$ space, denoted as $\mathscr{L}_{2e}$. More specifically, $\mathscr{L}_{2e}$ is the space of signals that have finite energy over all finite time horizons, but do not necessarily need to have finite energy in the infinite horizon. Formally, this signal space can be defined as $\mathscr{L}_{2e} \triangleq \left\{ a : \mathbb{R}_+ \to \mathbb{R}^n \mid \|a\|_{\mathscr{L}_{2[0,T]}} < \infty,\ \forall T < \infty \right\}$.

Given the above setup, as the start of our discussion, the classical metrics in robust control and fault detection are first examined, and we discuss to what extent they can or can not capture the attack scenario described in the previous section.

3 Classical metrics in robust control and fault detection

Typical worst-case metrics in control are the largest and smallest gains of a dynamical system. The largest gain is commonly used to capture the maximum amplification that the input can have on the output. On the contrary, the smallest gain captures the least amplification that the input has on the output. When the amplification is measured in terms of energy (*i.e.,* $\mathscr{L}_2$ signal norm), the largest and smallest gains are respectively the $\mathscr{H}_\infty$ norm and the $\mathscr{H}_-$ index.

3.1 The $\mathscr{H}_\infty$ norm

The $\mathscr{H}_\infty$ norm is a classical metric in robust control, capturing the largest energy amplification from input to output. In the context of our attack scenario, the adversary is interested in maximizing the energy of the performance output, y_c. Hence, we are interested in the worst-case (largest) amplification from the attack signal to the performance output, which is captured by the $\mathscr{H}_\infty$ norm of the system (5) from $a(t)$ to $y_c(t)$, namely $\|\Sigma_c\|_{\mathscr{H}_\infty}$.

There are multiple equivalent characterizations of the $\mathscr{H}_\infty$ norm. An appealing one for our purposes is the formulation as the following optimal control problem:

$$\begin{aligned} \|\Sigma_c\|^2_{\mathscr{H}_\infty} \triangleq \sup_{a \in \mathscr{L}_{2e}, x(0)=0} \quad & \|y_c\|^2_{\mathscr{L}_2} \\ \text{s.t.} \quad & \|a\|^2_{\mathscr{L}_2} \leq 1 \end{aligned} \tag{7}$$

Yet another useful interpretation of the $\mathscr{H}_\infty$ norm is that of the maximum $\mathscr{L}_2$ amplification of the system, from input to output, *i.e.*, $\|\Sigma_c\|^2_{\mathscr{H}_\infty} = \gamma \geq 0$ implies

$$\|y_c\|^2_{\mathscr{L}_2} \leq \gamma \|a\|^2_{\mathscr{L}_2}, \quad x(0) = 0. \tag{8}$$

Finally, defining $G_c(s) = C_c (sI - A)^{-1} B + D_c$, the $\mathscr{H}_\infty$ norm can also be related to the singular values of the system Σ_c:

$$\overline{\sigma}_c(w)^2 \leq \sup_{w>0} \overline{\sigma}_c(w)^2 = \gamma, \tag{9}$$

where

$$\overline{\sigma}_c(w)^2 \triangleq \max_{a \in \mathbb{C}^{n_a}} \frac{\|G_c(jw)a\|^2}{\|a\|_2^2}.$$

The $\mathscr{H}_\infty$ norm has well-known properties, for instance that the $\|\Sigma_c\|_{\mathscr{H}_\infty}$ is unbounded if and only if the system Σ_c is unstable. Moreover, note that the constraint in (7) essentially restricts the attack signal to have finite energy over infinite horizons. This in turns implies that the worst-case amplification does not consider attack signals with possibly infinite energy, namely non-vanishing signals, such as strictly increasing exponential signals and ramps.

Remark 1. Recalling Example 1 with the quadruple tank under attack on both actuators, we observe that the exponentially increasing, undetectable attack described in Example 1 is not considered by the $\mathscr{H}_\infty$ norm. Such a potentially harmful attack is therefore not included in analysis based on the $\mathscr{H}_\infty$ norm.

Furthermore, the $\mathscr{H}_\infty$ norm does not give any guarantees to the detectability of the attack signal. Hence, the worst-case attack signal may turn out to be easily detectable, despite resulting in the worst amplification to the performance output (with respect to other input signals with finite energy).

Example 2. Consider the quadruple tank system of Example 1, but now under attack on the first actuator (pump 1, corresponding to the first entry in the vector $u(t)$). Computing the $\mathscr{H}_\infty$ norm for such a scenario, and in particular examining the largest singular value $\overline{\sigma}_c(w)$, indicates that the worst-case frequency is $w = 0$. Hence, constant attacks will result in the worst-case amplification of the cost. For a constant attack of the form $a(t) = 1$, the worst-case in the $\mathscr{H}_\infty$ sense, the RMS of both the performance output and the detection output are shown in Fig. 3 (blue dotted line). Another attack signal at a higher frequency, but with the same level of impact on performance at steady-state, is also depicted (red dash-dotted line). As it can be observed from the lower plot, the new attack has a much lower level of detectability. From a security perspective, such an attack is thus of higher concern than the worst-case attack in the $\mathscr{H}_\infty$ sense.

As illustrated by the above discussions and the latter example, we need a metric that also accounts for detectability of the attack signal.

3.2 The $\mathscr{H}_-$ index:

As opposed to the $\mathscr{H}_\infty$ norm, the $\mathscr{H}_-$ index is a classical metric in fault detection, capturing the smallest energy amplification from input to output. In the context of our attack scenario, the adversary is interested in minimizing the energy of the

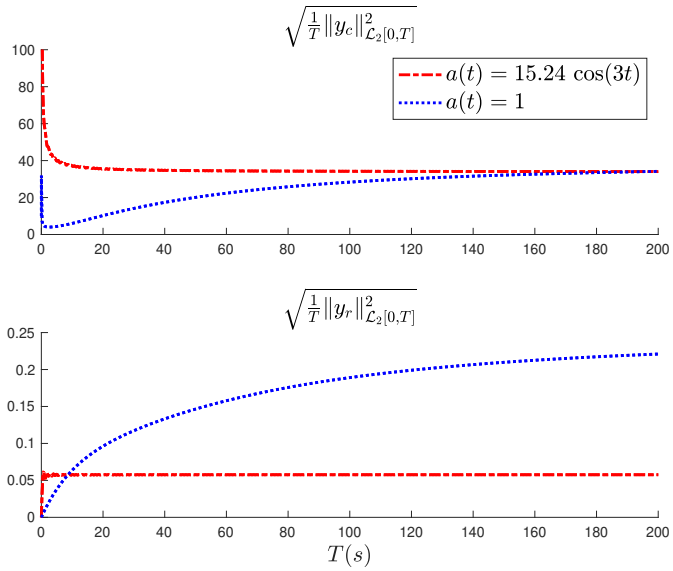

Fig. 3 The RMS values over time of the performance and detection outputs, y_c and y_r, for different attack signals on actuator 1. The worst-case attack, in the $\mathscr{H}_\infty$ sense, is depicted in dotted blue line. For the same level of impact on performance, another attack signal at a higher frequency (red dash-dotted line) is shown to have a much lower level of detectability.

detection output, y_r. Therefore, we are interested in the worst-case (smallest) amplification from the attack signal to the detection output, which is captured by the $\mathscr{H}_-$ index of the system (5) from $a(t)$ to $y_r(t)$, namely $\|\Sigma_r\|_{\mathscr{H}_-}$.

Similarly as for the $\mathscr{H}_\infty$ norm, there are multiple possible characterizations of the $\mathscr{H}_-$ index. First, we have again the formulation as the following optimal control problem:

$$\begin{aligned} \|\Sigma_r\|^2_{\mathscr{H}_-} \triangleq \inf_{a\in\mathscr{L}_{2e}, x(0)=0} \quad & \|y_r\|^2_{\mathscr{L}_2} \\ \text{s.t.} \quad & \|a\|^2_{\mathscr{L}_2} \geq 1. \end{aligned} \tag{10}$$

Second, a useful interpretation of the $\mathscr{H}_-$ index is that of the minimum $\mathscr{L}_2$ amplification of the system, *i.e.*, $\|\Sigma_r\|^2_{\mathscr{H}_-} = \gamma \geq 0$ implies

$$\|y_r\|^2_{\mathscr{L}_2} \geq \gamma \|a\|^2_{\mathscr{L}_2}, \quad x(0) = 0. \tag{11}$$

Finally, defining $G_r(s) = C_r(sI - A)^{-1}B + D_r$, the $\mathscr{H}_-$ index can also be related to the singular values of the system Σ_r:

$$\underline{\sigma}_r(w)^2 \geq \inf_{w>0} \underline{\sigma}_r(w)^2 = \gamma, \tag{12}$$

where

$$\underline{\sigma}_r(w)^2 \triangleq \min_{a\in\mathbb{C}^{n_a}} \frac{\|G_r(jw)a\|^2}{\|a\|_2^2}.$$

Remark 2. The original definition of the $\mathscr{H}_-$ in the fault detection literature (Liu et al, 2005) was based on the singular values (12). This definition is actually more conservative than the new formulations based on optimal control (10) and energy amplification (11). Specifically, the formulation in (12) implicitly constrains the input signal to have finite energy, and thus lie in $\mathscr{L}_2$. On the other hand, (10) and (11) allow the input signal to lie in lie in $\mathscr{L}_{2e}$ and thus have infinite energy in infinite time horizons.

The original $\mathscr{H}_-$ index defined in (12) is well-known for its limitation in strictly proper systems, in which case $\underline{\sigma}_r(w)$ continuously decreases for high frequencies, and thus $\mathscr{H}_- = 0$. Moreover, as per Remark 2, it again restricts the attack signal to have finite energy over infinite horizons. Hence, attacks that are typically dangerous, but hard to detect, such as incipient signals are not considered by the original $\mathscr{H}_-$ index.

The new formulations of the $\mathscr{H}_-$ index, namely (10) and (11) address some of the conservatism of (12), and consider, for instance, the presence of unstable zeros in the system that will render $\mathscr{H}_- = 0$, which was neglected in (12).

However, the $\mathscr{H}_-$ index does not give any measure to the impact of the attack signal on the closed-loop system performance. Hence, the worst-case attack signal may turn out to be hard to detect, but at the same time result in negligible impact on performance.

Example 3. Consider the quadruple tank system of Example 1, but now under attack on the second actuator (pump 2, corresponding to the second entry in the vector $u(t)$). Computing the $\mathscr{H}_-$ index for such a scenario, and in particular examining the smallest singular value $\underline{\sigma}_r(w)$, indicates that the detectability deteriorates monotonically with increasing frequencies. Thus the worst-case frequency is $w = \infty$, in which case $\|\Sigma_r\|_{\mathscr{H}_-} = 0$. Hence, attack signals with very high frequency will result in the worst-case detectability at the detection output.

The RMS of the performance and detection outputs for different attacks are illustrated in Fig. 4, where the attack magnitudes have been adjust so that all attacks have the same level of detectability at steady-state (*i.e.*, same RMS value for the detection outputs).

As seen from the magnitude of the different attacks, a large magnitude is required for obtaining the same level of detectability, which confirms that the detectability in the $\mathscr{H}_-$ sense decreases with frequency.

However, worse detectability alone does not imply a larger impact in performance. In fact, the attack with the least impact in performance (depicted by dash-dotted red plots) is less detectable than the constant attack (dotted blue plots).

Moreover, the impact on performance is not monotonic with frequency. Indeed, we observe that, for the same level of detectability, the most and the least detectable attacks (in dotted and dashed lines, respectively) yield the same RMS value for the performance output, and thus have the same level of impact.

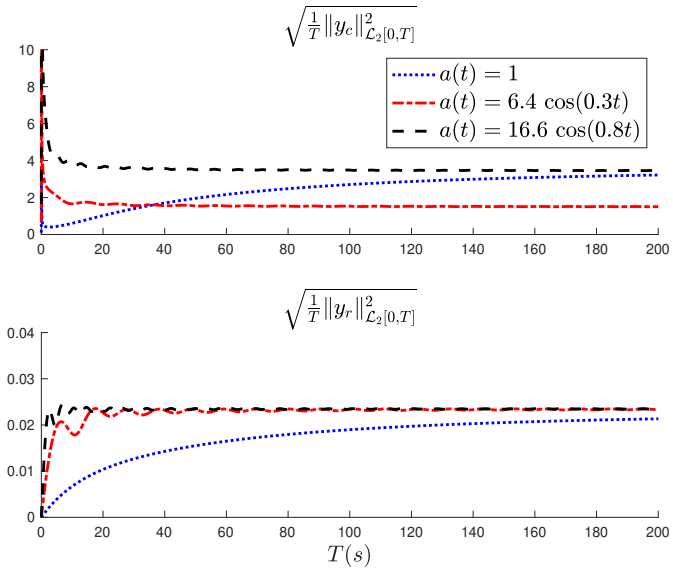

Fig. 4 The RMS values over time of the performance and detection outputs, y_c and y_r, for different attack signals on actuator 2. For the same level of detectability, the impact on performance does not decrease with the frequency.

3.3 Mixing $\mathscr{H}_\infty$ and $\mathscr{H}_-$

As illustrated in the previous subsection, the $\mathscr{H}_\infty$ norm and the $\mathscr{H}_-$ index cannot simultaneously capture both of the aspects that define malicious attacks: the aim to maximize impact on performance, while simultaneously minimizing detection.

Since $\mathscr{H}_\infty$ addresses the impact on performance, while $\mathscr{H}_-$ encodes detectability, a natural first approach would be to attempt to combine both metrics into one single quantity. However, as seen in Example 4, the worst-case frequency differs between the two metrics, which means that these metrics look into distinct worst-case input signals.

As a first attempt to combine the essence of these metrics, and develop a security metric, we consider instead the ratio between the singular values, $\mu(w)$, and define a first heuristic security metric as $\overline{\mu}$, with

$$\mu(w) \triangleq \frac{\overline{\sigma}_c(w)^2}{\underline{\sigma}_r(w)^2} \leq \sup_{w>0} \frac{\overline{\sigma}_c(w)^2}{\underline{\sigma}_r(w)^2} = \overline{\mu}, \tag{13}$$

There are, however, two main flaws in this approach. The first is that it constraints the attack signal to have finite energy, *i.e.* to lie in $\mathscr{L}_2$, and hence only considers vanishing signals. For instance, as for the $\mathscr{H}_-$ index, the presence of unstable zeros would not be considered, while it is well-known that attacks replicating

the unstable zero-dynamics are not detected and still have a dramatic impact on the system (Pasqualetti et al, 2013; Teixeira et al, 2015a).

The second is that, for multiple-input systems (e.g., when multiple sensors and actuators are corrupted), $\mu(w)$ does not consider the spatial coordination of the different attack channels to achieve a large impact and worse detection. In other words, the spatial coordination between attack channels is encoded separately in the definitions of $\overline{\sigma}_c(w)$ and $\underline{\sigma}_r(w)$, and thus each singular value will consider different worst-case attack direction (or singular vector).

Example 4. Consider the scenario in Example 3, where the quadruple tank system is under a cyber-attack on the second actuator. The singular values $\overline{\sigma}_c(w)$ and $\underline{\sigma}_r(w)$ are depicted in Fig. 5, as well as their ratio $\mu(w)$.

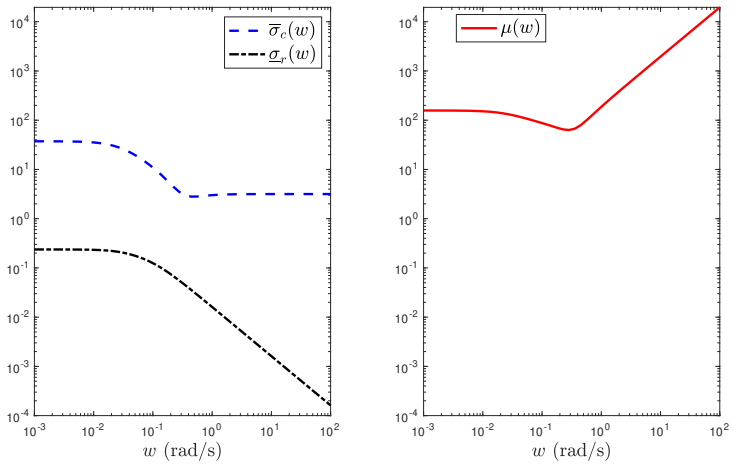

Fig. 5 The singular values $\overline{\sigma}_c(w)$ and $\underline{\sigma}_r(w)$ (left) and their ratio $\mu(w)$ (right) for a cyber-attack on actuator 2.

As discussed in Example 3 and observed from $\underline{\sigma}_r(w)$ in the figure, the detectability deteriorates monotonically with the frequency as $\|\Sigma_r\|_{\mathscr{H}_-}$ converges to 0. Hence, attack signals with very high frequency will result in the worst-case detectability at the detection output. This behavior is a natural consequence of the plant dynamics: in the absence of a direct feed-through term from $u(t)$ to $y_m(t)$, actuation signals with very high frequency are naturally blocked by the system dynamics, and will not appear in the measured output.

On the other hand, by observing $\overline{\sigma}_c(w)$ we conclude that the worst-case impact occurs at low frequencies. However, note that the $\overline{\sigma}_c(w)$ converges to a non-zero constant as the frequency increases. This occurs due to the nonzero feed-through term $D_c \neq 0$, which means that the attack on the actuator has a direct impact on the control cost (2).

Connecting the observations from the two classical metrics, we conclude that attack signals with high frequencies will be hard to detect, while having a non-zero impact on the system. The worst-case would indeed be to have an infinitely high frequency, in which case the attack would be completely undetectable by y_r, while still having a non-zero impact on y_c. This is in fact correctly captured by the mixed metric $\mu(w)$, which tends to infinity as the frequency increases.

The above example illustrates that the heuristic metric $\mu(w)$ can, for certain scenarios, correctly capture the worst-case malicious attacks. In particular, as only one signal was attacked in Example 4, there was no spatial coordination between attack signals to be considered. Next we briefly discuss an example where such spatial coordination is crucial.

Example 5. Consider the scenario where $E_p = B_p$, that is, the adversary is able to stage a physical attack through some additional pumps other than the plant's actuators, but with identical effect on the system's states. The singular values $\overline{\sigma}_c(w)$ and $\underline{\sigma}_r(w)$, as well as their ratio $\mu(w)$, are depicted in Fig. 6.

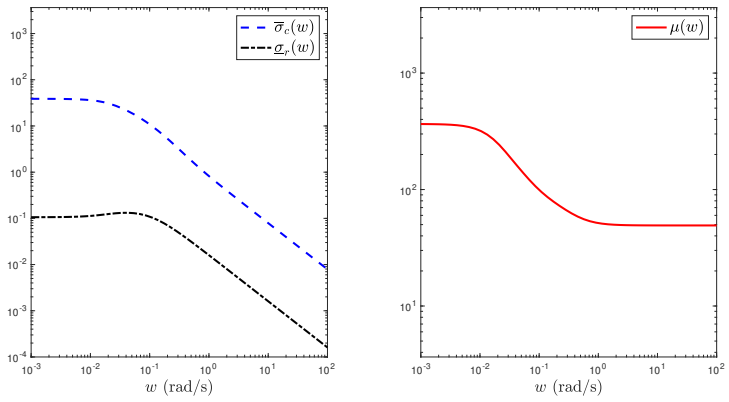

Fig. 6 The singular values $\overline{\sigma}_c(w)$ and $\underline{\sigma}_r(w)$ (left) and their ratio $\mu(w)$ (right) for a cyber-attack on actuator 2.

Since the attack does not affect the actual actuators, there is no direct feedthrough term from attack to any of the outputs (i.e., $D_c = D_r = 0$). Therefore, both the impact and the detectability decay to zero with the frequency. Moreover, it turns out that they decay at the same slope, which leads to a constant high frequency asymptote in $\mu(w)$. By observing $\mu(w)$, one would be led to conclude that the worst-case attack with maximum impact and minimum detectability happens for low frequencies.

On the contrary, there exists an attack that is undetectable and still leads to arbitrarily large impact. As discussed in Example 1, the plant has one unstable zero from the actuators to the measurement output, which can thus be exploited by the

physical attack, since $E_p = B_p$. Such attack is non-vanishing and requires the coordination of both water pumps, two aspects that are not captured by the heuristic metric $\mu(w)$.

As motivated by the above discussions and examples, the heuristic metric $\mu(w)$ correctly captures the impact of worst-case stealthy attacks is some scenarios (Example 4), but fails in other cases where spatial coordination is exploited by the adversary (Example 5). Thus, there is a need for an over-arching security metric that correctly tackles strategic stealthy attacks in general scenarios. One such security metric is described in the next section.

4 A security metric for analysis and design: the output-to-output gain

The previous section described the classical metrics in control and fault detection, and illustrated their shortcomings in capturing simultaneously the impact and the detectability of malicious attacks, which also carried over to the heuristic metric $\mu(w)$. In this section, a new security metric that can successfully consider these two aspects is discussed and characterized. Moreover, we show that such a metric can be the basis for designing controllers and detectors for increased resiliency against adversaries.

4.1 Security analysis with the output-to-output gain

A security metric simultaneously integrating the impact on performance and the detectability of attacks was proposed in Teixeira et al (2015a): the output-to-output gain (OOG). This metric is tailored to the malicious adversary objectives of maximizing impact while remaining undetected, being defined as the optimal control problem

$$\begin{aligned} \|\Sigma\|^2_{y_c \leftarrow y_r} \triangleq \sup_{a \in \mathscr{L}_{2e}, x(0)=0} \quad & \|y_c\|^2_{\mathscr{L}_2} \\ \text{s.t.} \quad & \|y_r\|^2_{\mathscr{L}_2} \leq 1. \end{aligned} \tag{14}$$

As before, a useful interpretation of the OOG is that of the maximum $\mathscr{L}_2$ amplification of the system, but now in terms of the two outputs $\|y_c\|^2_{\mathscr{L}_2}$ and $\|y_r\|^2_{\mathscr{L}_2}$, *i.e.*, $\|\Sigma\|^2_{y_c \leftarrow y_r} = \gamma \geq 0$ implies

$$\|y_c\|^2_{\mathscr{L}_2} \leq \gamma \|y_r\|^2_{\mathscr{L}_2}, \quad x(0) = 0. \tag{15}$$

A few remarks are in order. First, note that the attack signal is not *a priori* constrained to be a vanishing signal, as opposed to the $\mathscr{H}_\infty$ norm and the original $\mathscr{H}_-$

index. Thus, the OOG can consider unstable zeros, as well as exponentially increasing signals and ramps.

Second, the gain characterization is rather insightful. Recalling that no detection alarm is triggered while $\|y_r\|^2_{\mathscr{L}_2} \leq 1$, the condition (15) implies a guaranteed bound on the impact in performance for any attack that is not detected, namely, $\|y_c\|^2_{\mathscr{L}_2} \leq \gamma \|y_r\|^2_{\mathscr{L}_2} \leq \gamma$.

As such, the OOG seamlessly characterizes the worst-case impact on performance of undetected, possibly non-vanishing attacks.

Finally, we refer to a complete characterization of the OOG in terms of dissipative systems theory (Trentelman and Willems, 1991), as summarized in the following statement.

Proposition 1. *Consider the continuous-time system* $\Sigma \triangleq \left(A, B, \left[C_c^\top \, C_r^\top\right]^\top, \left[D_c^\top \, D_r^\top\right]^\top\right)$, *as described by* (5)*, which is assumed to be controllable. Define* $G_r(s) = C_r(sI - A)^{-1}B + D_r$ *and* $G_c(s) = C_c(sI-A)^{-1}B + D_c$.

The following statements are equivalent:

1. *The OOG of the system satisfies the bound* $\|\Sigma\|^2_{y_c \leftarrow y_r} \leq \gamma$*;*
2. *The system* Σ *is dissipative w.r.t. the supply rate* $s(x(t), a(t)) = \gamma \|y_r(t)\|_2^2 - \|y_c(t)\|_2^2$*;*
3. *For all trajectories of the system with* $T > 0$ *and* $x(0) = 0$*, we have*

$$\int_0^T s(x(t), a(t))\, dt \geq 0;$$

4. *There exists a* $P \succeq 0$ *such that the following Linear Matrix Inequality holds:*

$$R(\Sigma, P, \gamma) \triangleq \begin{bmatrix} A^\top P + PA & PB \\ B^\top P & 0 \end{bmatrix} - \gamma \begin{bmatrix} C_r \top \\ D_r^\top \end{bmatrix} \begin{bmatrix} C_r & D_r \end{bmatrix} + \begin{bmatrix} C_c \top \\ D_c^\top \end{bmatrix} \begin{bmatrix} C_c & D_c \end{bmatrix} \preceq 0. \quad (16)$$

Additionally, a necessary *condition of the previous statements to hold is that the following frequency-domain condition holds:*

$$\gamma G_r(\bar{s})^\top G_r(s) - G_c(\bar{s})^\top G_c(s) \succeq 0, \forall s \in \mathbb{C}, \quad (17)$$

with $s \notin \lambda(A)$, $Re(s) \geq 0$.

The above result follows directly from solving the optimal control problem (14) using dissipative systems theory (Willems, 1972), akin to the discrete-time case investigated in Teixeira et al (2015a), and recalling key results in dissipative system theory for linear systems with quadratic supply rates (Trentelman and Willems, 1991).

Note that statement 3 in Proposition 1 is precisely equivalent to the gain condition presented in (15). Moreover, and most importantly, the LMI in statement 4 leads to a computationally efficient approach to compute the OOG of a given system. In fact, the OOG can be obtained by solving the following convex optimization problem

$$\begin{aligned} \|\Sigma\|^2_{y_c \leftarrow y_r} = \min_{P \succeq 0, \gamma > 0} \quad & \gamma \\ \text{s.t.} \quad & R(\Sigma, P, \gamma) \preceq 0. \end{aligned} \tag{18}$$

Finally, the inequality in (17) provides a necessary frequency-domain condition for the OOG to be bounded. This is in contrast to similar frequency-domain conditions for the classical metrics in the previous subsections, which are both necessary and sufficient, and also significantly less involved than (17).

A necessary and sufficient frequency-domain condition was first derived in Molinari (1975). Unfortunately, this condition involves an infinite-dimensional inequality on the complex plane whose evaluation is not tractable (Trentelman, 1999). In pursuit of tractable conditions, under certain regularity assumptions, it was shown that (17) was also sufficient, see Gannot (2019, Section 2.3). Under milder regularity assumptions, a frequency-domain condition based on Pick matrices was also proposed in Trentelman and Rapisarda (2001).

Unfortunately, while the classical sensitivity metrics do satisfy the regularity assumptions investigated in the literature, the general formulation of the OOG does not. For instance, the case with $D_c = D_r = 0$ does not enjoy such regularity properties. Nonetheless, the necessary condition (17) points to useful structural results characterizing degenerate cases of the OOG, where the gain is unbounded regardless of the choice of controller and anomaly detector. Furthermore, it also provides a lower bound to the OOG, and it has a tight connection to the heuristic metric $\mu(w)$ defined in (13). These two aspects are further explored in the following.

4.1.1 Structural results on the output-to-output gain

The frequency-domain condition(17) is centered on the notion of unstable invariant zeros of a transfer function $G(s)$, that is, values $s \in \mathbb{C}$ (possibly at infinity) with $Re(s) \geq 0$ such that there exists $a \in \mathbb{C}^{n_a}$ for which $G(s)a = 0$. The precise result is as follows.

Proposition 2. *Consider the continuous-time system* $\Sigma \triangleq \left(A, B, \left[C_c^\top C_r^\top\right]^\top, \left[D_c^\top D_r^\top\right]^\top\right)$, *as described by* (5)*, which is assumed to be controllable. Define* $G_r(s) = C_r(sI - A)^{-1}B + D_r$ *and* $G_c(s) = C_c(sI-A)^{-1}B+D_c$. *The OOG gain of the system,* $\|\Sigma\|^2_{y_c \leftarrow y_r}$, *is bounded if, and only if, all the unstable invariant zeros of* $G_r(s)$ *(including zeros at infinity and their multiplicity) are also zeros of* $G_c(s)$.

A variation of this result has first appeared in Teixeira et al (2015a), for discrete-time systems. These conditions point to fundamental limitations in security under the presence of unstable zeros (including zeros at infinity), as observed in similar contexts in the literature (Pasqualetti et al, 2013; Mo and Sinopoli, 2012; Teixeira et al, 2012, 2015b). Such limitations are examined in the following example.

Example 6. Consider the scenarios in Examples 4 and 5. Computing the OOG for both scenarios, by solving the optimization problem (18), would yield an unbounded value for the gain. This result is in line with Proposition 2, as discussed below.

In Example 4, the multiplicity of the zeros at infinity of $G_r(s)$ is higher than that of $G_c(s)$, which leads to the increasing high-frequency asymptote of $\mu(w)$, and results in an unbounded OOG. As for Example 5, the system $G_r(s)$ has one unstable zero that is not a zero of $G_c(s)$ (since the direct term D_c is non-zero).

Since invariant zeros are not changed with output-feedback, one may conclude that the inherent zero structure of the open-loop system plays a crucial role in the sensitivity to stealthy attacks, regardless of the control and monitoring algorithms.

4.1.2 A lower bound on the output-to-output gain

Recalling the frequency-domain inequality (17), and inspired in (18), one can define the following variable:

$$\begin{aligned} \hat{\gamma} \triangleq \min_{\gamma>0} \quad & \gamma \\ \text{s.t.} \quad & \gamma G_r(\bar{s})^\top G_r(s) - G_c(\bar{s})^\top G_c(s) \succeq 0,\ \forall s \notin \lambda(A),\ Re(s) \geq 0. \end{aligned} \tag{19}$$

In the case where frequency-domain inequality (17) is both necessary and sufficient, the constraints (17) and the LMIs in statement 4 would be equivalent. Such an equivalence would transfer also to (18) and (19), which means that the OOG would be characterized as $\|\Sigma\|^2_{y_c \leftarrow y_r} = \hat{\gamma}$. However, as previously discussed, the frequency-domain inequality (17) is only necessary in general, and thus $\hat{\gamma}$ is generally only a lower bound of the OOG, i.e., $\|\Sigma\|^2_{y_c \leftarrow y_r} \geq \hat{\gamma}$.

Not surprisingly, the lower bound $\hat{\gamma}$ has a close connection to the singular values of the system. In fact, $\hat{\gamma}$ can be computed as

$$\hat{\gamma} = \sup_{s \in \mathscr{S}} \gamma(s),$$

where $\mathscr{S} \triangleq \{s \in \mathbb{C} :\ s \notin \lambda(A),\ Re(s) \geq 0\}$ and $\gamma(s)$ is defined as

$$\begin{aligned} \gamma(s) \triangleq \min_{\gamma \geq 0} \quad & \gamma \\ \text{s.t.} \quad & \gamma G_r(\bar{s})^\top G_r(s) - G_c(\bar{s})^\top G_c(s) \succeq 0. \end{aligned} \tag{20}$$

Note that $\gamma(s)$ essentially corresponds to the maximum generalized eigenvalue of the matrix pencil $(G_c(\bar{s})^\top G_c(s),\ G_r(\bar{s})^\top G_r(s))$, which may be interpreted as a generalized singular value of the system.

Finally, we highlight one interesting relation between the lower bound $\hat{\gamma}$ ($\gamma(s)$) and the heuristic $\overline{\mu}$ ($\mu(w)$). Consider the class of single-input systems, in which case $G_r(s)$ and $G_c(s)$ are complex-valued vector functions, denoted as $g_r(s) \in \mathbb{C}^{n_r}$ and $g_c(s) \in \mathbb{C}^{n_c}$ respectively. In such a case, the function $\gamma(s)$ can be rewritten as $\gamma(s) = \dfrac{\|g_c(s)\|_2^2}{\|g_r(s)\|_2^2}$. Observing that $\|g(s)\|_2^2 = \overline{\sigma}(s)^2 = \underline{\sigma}(s)^2$, one can re-write $\gamma(s)$ as

$\gamma(s) = \dfrac{\overline{\sigma}_c(s)^2}{\underline{\sigma}_r(s)^2}$, from which it follows that $\hat{\gamma}$ is bounded from below by $\overline{\mu}$, since

$$\hat{\gamma} = \sup_{s\in\mathscr{S}} \gamma(s) \geq \sup_{s\in\mathscr{S}\bigcap Re(s)=0} \gamma(s) = \sup_{w\notin\lambda(A)} \mu(w) = \overline{\mu}.$$

Hence, we conclude that, for single-input systems (i.e., where the adversary corrupts only one resource, such that $n_a = 1$), the heuristic metric $\mu(w)$ constructed in Section 3.3 can provide a lower bound to $\|\Sigma\|^2_{y_c\leftarrow y_r}$.

4.2 Security metrics-based design of controller and observer

The security metric described in the previous section, the output-to-output gain, bears strong similarities to the classical $\mathscr{H}_\infty$ norm and the $\mathscr{H}_-$ index. This points to the possibility of using the OOG to design controllers and anomaly detectors, as it happens with the classical metrics. In this section, we make a first exploration of a possible design for continuous-time systems, and discuss some of its properties.

Recall the closed loop system under attack Σ described by (5). Naturally, Σ depends on the actual choices of the observer and feedback gain matrices K and L, respectively. To highlight this dependency in this section, we use the notation $\Sigma(K,L)$.

From a design perspective, we wish to choose the matrices K and L that minimize the worst-case impact of attacks that are not detected. In summary, we look into approaches for choosing K and L such that the OOG gain of the corresponding system is minimized. Formally, the optimal K and L can be characterized as the optimal solutions to the following (non-convex) optimization problem

$$\begin{aligned} \min_{P\succeq 0,\gamma>0,K,L} \quad & \gamma \\ \text{s.t.} \quad & R(\Sigma(K,L),P,\gamma) \preceq 0, \end{aligned} \tag{21}$$

which follows directly from items 3 and 4 in Proposition 1.

Due to the products $A^\top P$ and $C_c^\top C_c$ in $R(\Sigma(K,L),P,\gamma)$ (cf. 4 in Proposition 1), the constraint in (21) is a Bilinear Matrix Inequality, which renders the optimization problem non-convex. Applying the Schur complement lemma allows us to remove the quadratic term $C_c^\top C_c$, obtaining instead the following problem

$$\begin{aligned} \min_{P\succeq 0,\beta>0,K,L} \quad & \beta \\ \text{s.t.} \quad & \begin{bmatrix} A^\top P + PA & PB & C_c^\top \\ B^\top P & 0 & D_c^\top \\ C_c & D_c & \beta I \end{bmatrix} - \beta \begin{bmatrix} C_r\top \\ D_r^\top \\ 0 \end{bmatrix} \begin{bmatrix} C_r & D_r & 0 \end{bmatrix} \preceq 0, \end{aligned} \tag{22}$$

where the optimal OOG is given by $\|\Sigma(K,L)\|_{y_c\leftarrow y_r} = \beta = \sqrt{\gamma}$. Given that C_r and D_r do not depend on K and L, the only cross terms between decision variable are now in $A^\top P + PA$ and $B^\top P$. Although the constraint is still a BMI, we can now search for

a sub-optimal solution through one of the different approaches to handle BMIs. For instance, in the following we propose a simple algorithm using the alternating minimization approach, where the decision variable tuples $\{K,L\}$ and $\{P\}$ are solved in alternating steps, with the other tuple fixed during each step.

Algorithm 1 Integrated OOG-based design of observer and feedback gain matrices.
Input: The data matrices describing (5): the system matrices A_p, B_p, C_m, C_p, D_p, and the adversary matrices Γ_u, Γ_y, E_p.
Auxiliary variables: A_k, B_k, $C_{c,k}$, $D_{c,k}$
Output: K^*, L^*, β^*, P^*

1: Set $k=0$, $P_{-1}=\infty$, $P_0=0$.
2: Find stabilizing K_0 and L_0.
3: **while** $||P_k - P_{k-1}|| \geq \varepsilon$ **do**
4: $A_k = A(K_k,L_k)$, $B_k = B(K_k,L_k)$, $C_{c,k} = C_c(K_k,L_k)$, and $D_{c,k} = D_c(K_k,L_k)$.
5:
$$(P_{k+1},\star) = \underset{P\succeq 0,\beta>0}{\arg\min} \quad \beta$$
$$\text{s.t.} \quad \begin{bmatrix} A_k^\top P + PA_k & PB_k & C_{c,k}^\top \\ B_k^\top P & 0 & D_{c,k}^\top \\ C_{c,k} & D_{c,k} & \beta I \end{bmatrix} - \beta \begin{bmatrix} C_r^\top \\ D_r^\top \\ 0 \end{bmatrix} \begin{bmatrix} C_r & D_r & 0 \end{bmatrix} \preceq 0.$$
6:
$$(K_{k+1},L_{k+1},\beta_{k+1}) = \underset{K,L,\beta>0}{\arg\min} \quad \beta$$
$$\text{s.t.} \quad \begin{bmatrix} A^\top P_{k+1} + P_{k+1}A & P_{k+1}B & C_c^\top \\ B^\top P_{k+1} & 0 & D_c^\top \\ C_c & D_c & \beta I \end{bmatrix} - \beta \begin{bmatrix} C_r^\top \\ D_r^\top \\ 0 \end{bmatrix} \begin{bmatrix} C_r & D_r & 0 \end{bmatrix} \preceq 0.$$
{The matrices A, B, C_c, and D_c depend on K and L, as detailed in (6).}
7: $k = k+1$.
8: **end while**
9: **return** $(K^*,L^*,\beta^*,P^*) = (K_k,L_k,\beta_k,P_k)$.

Next we discuss an example where Algorithm 1 is used to re-design the closed-loop system.

Example 7. Consider a scenario where the quadruple tank system is subject to a physical attack directly on the first tank (*i.e.*, the first entry of the state vector $x_p(t)$). This attack scenario can be modeled as (5) with $\Gamma_u = \Gamma_y = \emptyset$ and $E_p = e_1$, which leads to a single-input system with $n_a = 1$. Therefore, the results from Section 4.1.2 hold, and the heuristic $\mu(w)$ provides a lower bound to the OOG, $\|\Sigma\|_{y_c \leftarrow y_r}$.

For illustration purposes, the singular values $\overline{\sigma}_c(w)$ and $\underline{\sigma}_r(w)$, and their ratio $\mu(w)$, are depicted in Fig. 7. The OOG of the system, computed through (18), is $\|\Sigma\|_{y_c \leftarrow y_r} = 36.4$. This value is in accordance with the peak of $\mu(w)$, that is, the OGG and $\overline{\mu} = \sup \mu(w)$ coincide, which indicates that the lower bound results from Section 4.1.2 are tight in this case.

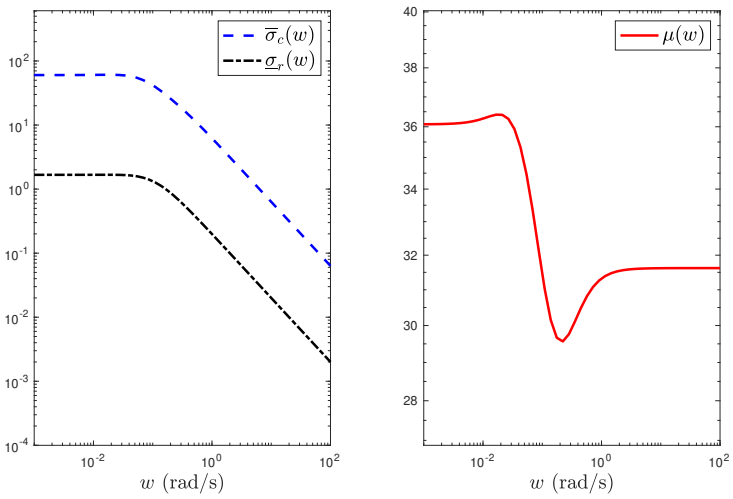

Fig. 7 Original closed-loop system. The singular values $\overline{\sigma}_c(w)$ and $\underline{\sigma}_r(w)$ (left) and their ratio $\mu(w)$ (right) for a physical attack on water tank 1.

Next, we leverage Algorithm 1 to design an improved observer-based controller and detector. Applying the algorithm to the closed-loop system under attack yields the following gain matrices

$$K = \begin{bmatrix} -0.2945 & 0.3364 \\ -0.4852 & -1.1088 \\ 0.0586 & 0.1256 \\ 0.0218 & -0.2697 \end{bmatrix}, \quad L = \begin{bmatrix} -0.5130 & 0.0280 & 0.1759 & -0.5569 \\ 0.0183 & -0.3563 & -0.8769 & 0.1434 \end{bmatrix}.$$

The resulting singular values $\overline{\sigma}_c(w)$ and $\underline{\sigma}_r(w)$, and their ratio $\mu(w)$, are shown in Fig. 8. As before, the OOG of the system computed through (18) and $\sup \mu(w)$ coincide and are equal to $\|\Sigma\|_{y_c \leftarrow y_r} = 31.6$.

Several remarks are in order. First, note that the shape of the ratio $\mu(w)$ flipped after the design, indicating that the worst-case inputs moved from low frequencies (before the re-design) to the high frequencies (after the re-design).

Second, we observe that the performance singular value $\overline{\sigma}_c(w)$ became larger at the low frequencies after the re-design. Such an increase mean that the $\mathscr{H}_\infty$ norm of the system has increased after re-design, which may initially be counter-intuitive.

A third observation points to the possible usefulness of increasing the system's $\mathscr{H}_\infty$ norm: it has allowed the detectability singular value to increase substantially at low frequencies. This in turn implies that low frequency attacks became much more detectable (by a factor of 10). Therefore, although the impact has slightly increased at low frequencies, the detectability has greatly increased as well. This effect is

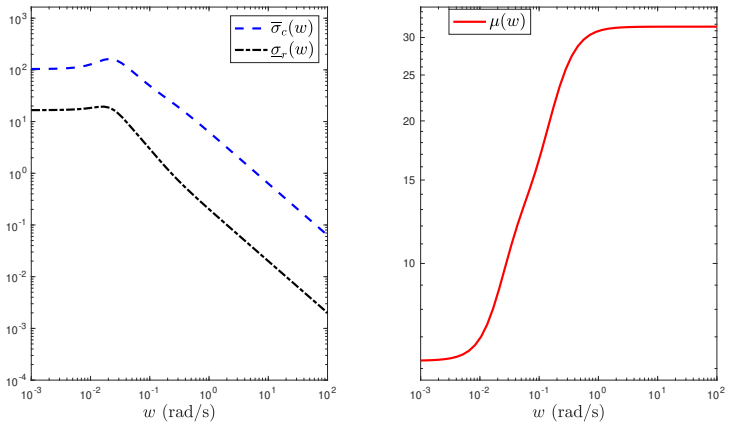

Fig. 8 Closed-loop system re-designed through Algorithm 1. The singular values $\overline{\sigma}_c(w)$ and $\underline{\sigma}_r(w)$ (left) and their ratio $\mu(w)$ (right) for a physical attack on water tank 1.

clearly visible by comparing the low frequency asymptotes of $\mu(w)$ before and after the re-design.

It is noteworthy to highlight that such trade-offs between impact and detectability, naturally imbued in the design procedure, are currently unavailable through existing techniques in robust control and fault detection. In particular, existing $\mathscr{H}_-$-based approaches would unlikely change low frequency detectability, since the worst-case detection in the $\mathscr{H}_-$ index sense occurs at very large frequencies.

5 Conclusions

In this chapter, we have considered the security of control systems, in scenarios where malicious adversaries aim at maximizing the impact on control performance, while simultaneously remaining undetected. The objectives of the chapter were to investigate possible metrics to analyze, and re-design, the closed-loop system from a security perspective.

Classical metrics from robust control and fault detection were revisited under the context of malicious attacks. The conclusion was that these metrics consider separately impact of attacks and detection, and are thus inadequate for security analysis. A initial attempt to merge these metrics was taken, by formulating a new metric consisting of the ratio between performance and detection singular values. Unfortunately, such an approach kept the limitations of classical metrics, and did not fully capture some of the known potentially dangerous attacks, that have arbitrarily large impact and low detectability.

A recently proposed security metric, the output-to-output gain (OOG), was then introduced and characterized. Borrowing results from dissipative systems theory for linear systems with quadratic supply rates, the OOG was entirely characterized. This in turn led to results enabling its efficient computation through convex optimization problems. Necessary and sufficient conditions describing fundamental limitations of the OOG were thus described.

Additionally, a first step was taken to use the OOG as a basis for controller and detector design. The OOG-based design problem was cast as a non-convex optimization problem with BMI constraints. Using the heuristic of alternating minimization to address the BMI constraints, an algorithm was proposed that results in a sub-optimal closed-loop system minimizing the OOG.

The results and insights contained in the chapter were supported and illustrated through several numerical examples on a common closed-loop system, which were presented continuously throughout the chapter.

Acknowledgements This work is financed by the Swedish Foundation for Strategic Research, and by the Swedish Research Council under the grant 2018-04396.

References

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

Cárdenas AA, Amin S, Sastry SS (2008) Secure control: Towards survivable cyber-physical systems. In: First Int. Work. Cyber-Physical Syst.

Fawzi H, Tabuada P, Diggavi S (2014) Secure estimation and control for cyber-physical systems under adversarial attacks. IEEE Trans on Autom Control 59(6):1454–1467

Gannot O (2019) Frequency criteria for exponential stability. arXiv:191010855 [math] 1910.10855

Johansson KH, Horch A, Wijk O, Hansson A (1999) Teaching multivariable control using the quadruple-tank process. In: 38th IEEE Conf. Decis. Control, pp 807–812

Liu J, Wang JL, Yang GH (2005) An LMI approach to minimum sensitivity analysis with application to fault detection. Automatica 41(11):1995–2004

Mo Y, Sinopoli B (2012) Integrity attacks on cyber-physical systems. In: 1st Int. Conf. High Confid. Networked Syst. CPSWeek 2012

Mo Y, Sinopoli B (2016) On the performance degradation of cyber-physical systems under stealthy integrity attacks. IEEE Trans Autom Control 61(9):2618–2624

Molinari B (1975) Conditions for nonpositive solutions of the linear matrix inequality. IEEE Trans Autom Control 20(6):804–806

Pasqualetti F, Dorfler F, Bullo F (2013) Attack detection and identification in cyber-physical systems. IEEE Trans Autom Control 58(11):2715–2729

Scherer C, Weiland S (2010) Linear matrix inequalities in control. In: Levine WS (ed) Control Syst. Handb. Control Syst. Adv. Methods, CRC Press

Shames I, Farokhi F, Summers TH (2017) Security analysis of cyber-physical systems using H2 norm. IET Control Theory Appl 11(11):1749–1755

Smith RS (2015) Covert misappropriation of networked control systems: Presenting a feedback structure. IEEE Control Syst 35(1):82–92

Teixeira A, Shames I, Sandberg H, Johansson KH (2012) Revealing stealthy attacks in control systems. In: 50th Annu. Allert. Conf. Commun. Control. Comput.

Teixeira A, Sandberg H, Johansson KH (2015a) Strategic stealthy attacks: The output-to-output l2-gain. In: Proc. IEEE Conf. Decis. Control, vol 54rd IEEE, pp 2582–2587

Teixeira A, Shames I, Sandberg H, Johansson KH (2015b) A secure control framework for resource-limited adversaries. Automatica 51(1):135–148, DOI 10.1016/j.automatica.2014.10.067, 1212.0226

Teixeira A, Sou K, Sandberg H, Johansson K (2015c) Secure control systems: A quantitative risk management approach. IEEE Control Syst Mag 35(1):24–45

Trentelman HL (1999) When does the algebraic Riccati equation have a negative semi-definite solution? In: Blondel V, Sontag ED, Vidyasagar M, Willems JC (eds) Open Problems in Mathematical Systems and Control Theory, Communications and Control Engineering, Springer, London, pp 229–237

Trentelman HL, Rapisarda P (2001) Pick Matrix Conditions for Sign-Definite Solutions of the Algebraic Riccati Equation. SIAM J Control Optim 40(3):969–991

Trentelman HL, Willems JCC (1991) The dissipation inequality and the algebraic Riccati equation. In: Bittanti S, Laub AJ, Willems JC (eds) Riccati Equ., Communications and Control Engineering Series, Springer Berlin Heidelberg, pp 197–242

Umsonst D, Sandberg H, Cardenas AA (2017) Security analysis of control system anomaly detectors. In: 2017 Am. Control Conf., IEEE, pp 5500–5506

Wang JL, Yang GH, Liu J (2007) An LMI approach to H_- index and mixed H_-/H_∞ fault detection observer design. Automatica 43(9):1656–1665

Willems JC (1972) Dissipative dynamical systems part II: Linear systems with quadratic supply rates. Arch Ration Mech Anal 45(5):352–393

Zhou K, Doyle JC, Glover K (1996) Robust and Optimal Control. Prentice-Hall, Inc., Upper Saddle River, NJ, USA