공학박사학위논문

Design and Countermeasure of Stealthy Attacks on Cyber-Physical Systems in Sampled-data Framework

샘플 데이터로 표현되는 사이버-물리 시스템의 취약점 분석 및 검출 불가능한 공격에 대한 방어 기법

2020년 8월

서울대학교 대학원

전기정보공학부

김 지 한

ABSTRACT

Design And Countermeasure of Stealthy Attacks On Cyber-Physical Systems in Sampled-Data Framework

by

Jihan Kim

Department of Electrical and Computer Engineering
College of Engineering
Seoul National University

August 2020

The rapid evolution of communication network and computation speed has led to the emergence of cyber-physical systems in which the traditional physical plants are controlled remotely using digital controllers. Unfortunately, however, the separation between the plant and controller with a network communication provides a new chance for external adversaries to intrude control systems, which are highly connected to human life and social infrastructures. For this reason, among various issues of the cyber-physical system, security problems have gained particular attention to control engineers these days. This dissertation presents new theoretical vulnerabilities undetectable from the conventional anomaly detector, which arise due to the mixture of continuous- and discrete-time components on cyber-physical systems, and addresses countermeasures against such vulnerabilities. Specific subjects dealt with in the dissertation are listed as follows:

- Zero dynamics attacks can be lethal to cyber-physical systems because they can be harmful to physical plants and impossible to detect. Fortunately, if the given continuous-time physical system is minimum phase, the attack is not so effective even if it cannot be detected. However, the situation can become unfavorable if one uses digital control by sampling the sensor measurement and using a zero-order hold for actuation because of the 'sampling zeros.' When the continuous-time system has a relative degree greater than two and the sampling period is small, the sampled-data system must have unstable zeros, so that the cyber-physical system becomes vulnerable to 'sampling zero dynamics attack.' In this dissertation, we present an idea to neutralize the zero dynamics attack for single-input and single-output sampled-data systems by shifting the unstable discrete-time zeros into stable ones. This idea is realized by employing the so-called 'generalized hold' which replaces a standard zero-order hold. It is shown that, under mild assumptions, a generalized hold exists which places the discrete-time zeros at desired positions. Furthermore, we formulate the design problem as an optimization problem whose performance index is related to the inter-sample behavior of the physical plant, and propose an optimal gain which alleviates the performance degradation caused by generalized hold as much as possible, and in order to verify the theoretical results, we apply the proposed strategy to a DC/DC converter with an electrical circuit.

- The zero dynamics attack has usually been studied as a type of actuator attack, but it can harm the physical plant through the sensor network. Specifically, when the system monitors abnormal behavior of the plant using the anomaly detector (fault detector), one can generate zero dynamics attack on the sensor network deceiving the anomaly detector by regarding the output of the plant and residual of the anomaly detector as a new input and output of a target system. It is noticed that this sensor attack is not so effective when the plant is stable even if the attack is still undetectable. Noting this point, we propose to reexamine the generalized hold as a countermeasure against the undetectable sensor attack. That is, using the fact that the output feedback passing through the generalized hold can stabilize the unstable

systems by selecting an appropriate hold function, we show that the plant can be safe from the undetectable sensor attack. Furthermore, to relieve the performance degradation of the use of generalized hold feedback, we employ a discrete-time linear quadratic regulator minimizing a continuous-time cost function.

- In the sampled-data framework, most anomaly detectors monitor the plant's output only at discrete time instants. Consequently, abnormal behavior between sampling instants cannot be detected if output behaves normally at every sampling instant. This implies that if an actuator attack drives the plant's state to pass through the kernel of the output matrix at each sensing time, then the attack compromises the system while remaining stealthy. This type of attack is always constructible when the sampled-data system has an input redundancy, i.e., the number of inputs being larger than that of outputs and/or the sampling rate of the actuators being higher than that of the sensors. Simulation results for the X-38 vehicle and other numerical examples illustrate this new attack strategy may result in disastrous consequences.

Keywords: cyber-physical system, network control system, sampled-data system, zero-dynamics attack, generalized hold
Student Number: 2014–21659

응원과 조언의 말씀을 주신 모든 분들께 감사드립니다.

Contents

List of Figures

Symbols and Acronyms

$\mathbb{Z}$	set of integers
$\mathbb{N}$	set of positive integers
$\mathbb{Q}$	set of rational numbers
$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers
$\lfloor \cdot \rfloor$	(component-wise) floor function
$\otimes$	Kronecker product
$\square$	end of theorems, lemmas, propositions, assumptions, remarks, and corollaries
$\blacksquare$	end of proof
X^n	set of n-tuples of elements in X
$X^{m \times n}$	set of m by n matrices of elements in X
$X(w)$	Fourier transform of $x(t)$
$\| \cdot \|_\infty$	infinity norm of a matrix or a vector
$\| \cdot \|$	Euclidean norm of a matrix or a vector
$\| \cdot \|_1$	1-norm of a matrix or a vector
$\| \cdot \|_M$	Maximum norm of a matrix; i.e., $\|A\|_M := \max_{i,j} \|a_{ij}\|$
$\sup(\cdot)$	supremum of a subset of $\mathbb{R}$, or a real-valued sequence
$\inf(\cdot)$	infimum of a subset of $\mathbb{R}$, or a real-valued sequence

$\cup$	union of sets
$\top$	transpose operator
$\mathrm{col}(y_1, y_2, \ldots, y_p)$	$:= [y_1^\top, y_2^\top, \cdots, y_p^\top]^\top$
$\sum_{i=0}^{n} b_i$	$:= b_0 + b_1 + \cdots + b_n$
T^{-1}	inverse matrix of square matrix T
$\det(T)$	determinant of square matrix T
$\mathrm{rank}(T)$	rank of matrix T
I_n	$n \times n$ identity matrix
$0_{m\times n}$	$m \times n$ zero matrix
$\frac{\partial \gamma}{\partial \mathsf{r}}$	partial derivative of function γ with respect to r
$\|\mathcal{X}\|$	$:= \sup(\{\|v\| : v \in \mathcal{X}\})$
$\exp(T)$	exponential of matrix T
$\int_a^b T(\tau)\, d\tau$	definite integral of function T on the set $\{\tau \in \mathbb{R} : a \leq \tau \leq b\}$
$\forall$	for all

- Given matrices $A \in \mathbb{R}^{n\times n}$ and $C \in \mathbb{R}^{m\times n}$, the pair (A, C) is said observable, when $\mathrm{rank}([C^\top, A^\top C^\top, \cdots, (A^{n-1})^\top C^\top]^\top) = n$.

- A matrix $H \in \mathbb{R}^{n\times n}$ is Hurwitz when all eigenvalues of H have negative real parts.

- By $a(t) \equiv a^*$ where $a^* \in \mathbb{R}$, it means the given signal $a(t)$ satisfies $a(t) = a^*$ for all $t = 0, 1, \cdots$.

- A set $X \subset \mathbb{R}^n$ is said compact, when it is closed and bounded.

- A characteristic polynomial of a square matrix $T \in \mathbb{R}^{n\times n}$ is defined as $\det(sI_n - T)$, as a polynomial with indeterminate s.

Acronyms

LTI	Linear time-invariant
CT	Continuous-time
DT	Discrete-time
SISO	Single-input single-output
MIMO	Multi-input Multi-output
DoS	Denial-of-Service
ZOH	Zero-order hold
GH	Generalized hold
CPS	Cyber-physical system
LQR	Linear quadratic regulator
DLQR	Discrete-time linear quadratic regulator
LPF	Low-pass filter
PWM	Pulse width modulation
MCU	Micro control unit

Chapter 1

Introduction

1.1 Overview of Security Issues on Cyber-Physical Systems

In traditional control systems, various control objects and controllers are connected using wired or local wireless communication because they are physically located close together. This paradigm has been changed in modern control systems as the rapid evolution of computing power of the electrical devices and data transmission speed of the network communication. Now, there is no need to keep the physical plant and digital controller close together, and which has led to the emergence of cyber-physical systems and network control systems. Many industries where it is difficult for a person to actually get close to a physical plant, have started to employ this cyber-physical system (or network control system), and it has been expanded to various control systems such as smart vehicles, drones, smart grids, power plants, etc.

Unfortunately, however, along with the advantages of the cyber-physical system, a new challenge came up into the picture to control engineers. The separation between the physical plant and controller with a network communication provides a new opportunity for external adversaries who want to intrude or hack control systems (Figure 1.1). The serious point is that cyber-physical systems are highly connected to human society and social infrastructures by its nature, which means cyber-attacks on cyber-physical systems may cause disastrous damage to human life and cost on critical infrastructures. These concerns are not fictitious ones. In

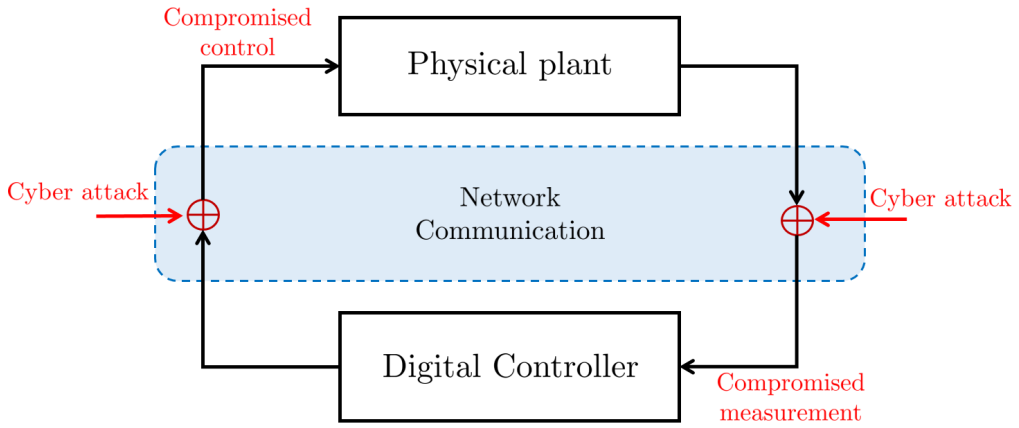


Figure 1.1: General concept of cyber attack on cyber-physical systems

fact, there have been many reports on real threats to cyber-physical systems. For instance, the famous Stuxnet attack on Iran's nuclear plant [Lan11, Kar11], the incident that Iran hacked U.S.'s unmanned aerial vehicle RQ-170 [Sta11], a massive power blackout in South America [Con10], SCADA security incident at Maroochy water service in Queensland, Australia [SM07], a cyber attack on SCADA of Ukrainian power plant [LAC16], and many more cyber security incidents can be found in [AMAY$^+$18] and references therein.

Due to the emergence of those real threats, among various issues of the cyber-physical system, the security problems have gained particular attention to control engineers. As a result, vast literature dealing with various types of cyber attacks and countermeasures has emerged in control-theoretic perspective. This trend of research seems like a duel between spears and shields; that is, when an idea of attack is devised in terms of an adversary, another idea to protect the system from it is developed in terms of the defender. As examples of this trend, some of the typical security issues of the cyber-physical system are listed below.

- **Replay attack:** replay attack hijacks and records information passing through the sensor network, and then reinjects the recorded data into the sensor network. This makes the replay attack be undetectable even if a malicious attack signal has injected into the input network to compromise the physical plant. Subsequently, when this attack became known to the

public, the idea of injecting watermarking signal into the control input is presented to identify the replay attack [MS09, KKS17, FQCZ20].

- **False data injection attack:** the false data injection attack showed that if an adversary can access to the output measurement of the in the electric power grid system and can inject an attack signal constructed with a linear combination of the output matrix of the system, then it can deceive the pre-installed state estimator [MS10, LNR11]. After that as a countermeasure, under the assumption of redundant observability, the resilient state estimation technique has been actively studied that can identify the compromised sensors based on the majority vote rule [FTD14, LSE15, LSE19].

- **DoS (Denial-of-Service) attack:** In cyber-physical systems, control signals and measurements are transmitted over communication networks. Hence, if malicious adversaries may have an authority to access those networks, they can jam or compromise the signals, and it causes delay, ruin, or loss of the packets on the communication network, and we call this a DoS attack. In order to overcome the DoS attack, control engineers employed optimal control techniques that make the control system preserve a minimum level of performance even when some packets are not delivered to the controller or physical plant [ACS09, ZCSC16, SPY$^+$17].

- **Zero dynamics attack:** zeros of the system have been regarded as a critical component since output-zeroing and state-diverging attack signals always exist when the plant has at least one unstable zero. This fact leads to the invention of the zero dynamics attack, which is lethal for general non-minimum phase systems due to its inherent stealthiness. Due to the zero dynamics attack utilizes specific information regarding zeros of the system, in order to prevent zero dynamics attack, the control engineers introduced ideas of altering the system configuration, inserting concealed information or time-varying component to change the zero's locations [TSSJ12, WS15, HZ16, WLS17, MA18].

In the meantime, as researches on various attack scenarios were conducted, literature has also emerged that specifies cyber-attacks according to resources

that the malicious adversary can utilize [TSSJ15]. The resources are composed to disruptive, disclosure, and model knowledge, and each of which describes the capability of the adversary that can access into the input or output channel to inject malicious attack, capture the information from the input or output channel, and have model knowledge of the whole system, respectively. This research is meaningful because it proposed the systematic categorization of various attacks.

1.2 Contributions and Outline of Dissertation

Among the various types of attacks, in this dissertation, we are mainly interested in analyzing and designing undetectable attacks. In particular, we focus on the attacks that are constructed with full model knowledge of the targeted system because, in terms of adversaries, it is possible to construct attacks more delicate way by exploiting full model knowledge, and which suggests more challenging problems to both attackers and defenders. In the following, Chapters 2 and 3 focus on examining the zero dynamics attack on the input and the output of the sampled-data system, respectively, and then propose new countermeasures to neutralize zero dynamics attack. In Chapter 4, we present a new possible sophisticated attack design mechanism to the multi-rate sampled-data system where the operation times of the holding device and sampler are allowed to be different.

Chapter 2. Use of Generalized Hold in the Sampled-data System to Counteract Against Zero Dynamics Attack

The zero dynamics attack [TSSJ12] is one of the most notorious threats because it is fundamentally stealthy while it is lethal for any non-minimum phase system. Earlier ideas of zero dynamics attack are heavily relying on the knowledge of exact system model. In particular, the attacker needs to know the exact location of the zeros. Inspired by this weakness, [HZ16] proposed to multiply a modulation matrix to the input stage in order to shift the location of zeros to somewhere unknown to the adversaries. This should be done secretly because the attacker can also counteract the modulation when the modification is known to the attacker. In terms of the attackers, they can also use the robust zero dynamics attack [PSL$^+$16, PLS$^+$19] which does not require the exact knowledge

of the zero locations, as long as the system is non-minimum phase. Does this mean that the system is safe if it is minimum phase? Unfortunately, even if the physical system that is given in the continuous-time domain is minimum phase, its sampled-data description may have unstable zeros, and this always happens when the continuous-time system has relative degree greater than two and the sampling is performed sufficiently fast [YG14]. Since almost all control systems work in the sampled-data domain, this is somehow inevitable in modern control systems[1].

Recently, instead of developing shields customized to particular spears as above, somewhat fundamental way is developed to protect the system from the zero dynamics attack. The idea of employing multi-rate sensing by [NHV15, NHV19] is to enforce the sampled-data system not to have unstable zeros. Then, since all the zeros of the sampled-data system are stable, there is no motivation for the adversary to engage the zero dynamics attack. Indeed, the zero dynamics attack may still remain stealthy, but its effect becomes negligible. However, as will be seen in Section 2.1.4, this method becomes less effective under fast sampling. Therefore, inheriting the same philosophy of [NHV15, NHV19], we propose a dual solution for single-input and single-output (SISO) sampled-data systems, which also enforces the sampled-data system to become minimum phase by moving or changing the zeros of the system. This is actually achieved by replacing the zero-order hold (ZOH) unit in the actuator by a generalized hold (GH) to be proposed. The idea of using GH to change the zeros in the sampled-data system is not new (see [Kab87]), but we propose its use as a countermeasure against the zero dynamics attack [BKL$^+$17]. Moreover, we present a way to design the GH through an optimization problem aiming to alleviate the side-effect of replacing the ZOH. Specifically, the solution to the optimization problem aims to minimize the difference of the hold gains between the GH and the ZOH while the discrete-time zeros of the resulting system are located inside the unit circle in the complex plane [KBP$^+$20]. Furthermore, we show that this problem can be cast into a convex problem so that the solution can be easily obtained by using popular solvers (e.g., `cvxgen` [MB12]). Subsequently, in order to verify the proposed theoretical

[1] In Section 2.1.2, we demonstrate this 'sampling zero dynamics attack'.

results, we carry out an experiment with a DC/DC converter connected to high-order low pass filter. In the last, a study regarding how can the generalized hold effect on intrinsic zeros of nonlinear systems is introduced.

Chapter 3. Use of Generalized Hold Feedback in the Sampled-data System to Counteract Zero Dynamics Sensor Attack with DLQR

In Chapter 3, we focus on the zero dynamics attack on the sensor, which makes the system state diverge with being stealthy to the anomaly detector that is the most common technique for detecting abnormal behaviors. Throughout the paper, we call this sensor attack as "*zero dynamics sensor attack.*"

The idea for constructing the zero dynamics sensor attack is the same that of the zero dynamics attack on the actuator except that the target system is altered from the physical plant to a composite system consisting of the state estimator and the anomaly detector. However, a notable difference occurs between the actuator and the sensor attack; that is, the latter is effective when the plant is unstable, whereas the former is activated properly when the plant is non-minimum phase.

Noting the fact that the zero dynamics sensor attack is effective for unstable systems, we propose a solution that is to move the unstable poles of the system into stable ones by employing an output feedback loop composed of the generalized hold device [YG14]. In fact, the idea that the poles of the sampled-data system can be shifted by implementing output feedback using generalized holder is not new [Kab87, HFA90, FG96], yet we propose to employ it for neutralizing the zero dynamics sensor attack, and we call this output feedback loop as "*generalized hold feedback.*" On the other hand, the use of the generalized hold feedback can lead to unfavorable fluctuations in the inter-sample behavior of the continuous-time plant. To relieve such an additional problem, we use a discrete-time linear quadratic regulator that minimizes the continuous-time performance index, not the discrete-time one.

Chapter 4. A New Vulnerability in the Multi-rate Sampled-data System: Design of Masking Attack

Apart from the system zero-based approaches, in this chapter, we are interested in investigating another type of system property which can be employed to design a fatal cyber-attack when exploited by the attacker. This new prop-

erty we are concerned with is an *input redundancy*, which exists when the target system has a sort of freedom in the input channel compared to the output. Roughly speaking, when the target system has the input redundancy, the way of constructing an attack signal can be separated into two parts; first, an attack signal is selected to enter the system through the redundant inputs and to enforce the system states diverge; next, a secondary signal is added to conceal or mask the influence of the diverging state to the output measurements. For this reason, we call such attack as *masking attack*. It is readily expected that, due to the nature of the attack, the adversary with the masking attack could enjoy the full advantage of other lethal attacks including the zero dynamics attack (i.e., the stealthiness and the ability to disrupt the system).

While a rough idea of masking attack was briefly mentioned in a recent work [NHV15], we newly propose a unified approach of the masking attack design for a general class of CPS in the sampled-data framework. The sampled-data systems of our interest are assumed to consist of a multi-input multi-output plant in continuous time, and a sampler and a zero-order holder (ZOH) with possibly different sampling periods (i.e., multi-rate sampling that has been studied for various purposes [SW00, HA02, FH02, FKK03, lSAQ07, KU08, MTO07]). In the general setting, the basic idea for the attack construction is twofold: first, the sampled-data system is represented as an extended *lifted system* with stacked input and output variables; then, with a particular condition on the output matrix of the lifted system, we select a bundle of attack signals at once, which can be arbitrarily large but enforce the state to remain in the (nontrivial) kernel of its output matrix at every sampling time of the output measurement. Specifically, this work shows that the requirement on the output matrix is satisfied when the input of CPS is redundant in the following two senses: (a) the sampling rate of the actuator is faster than that of the sensor, or (b) the number of inputs is larger than that of outputs. It is noteworthy that a sampled-data system equipped with this input redundancy may not have unstable zeros; therefore, the proposed masking attack could be yet another threat to CPS that are expected to be resilient against the widely-known stealthy attacks including the zero dynamics attack.

The present work allows the system class that the ratio of the sampling period

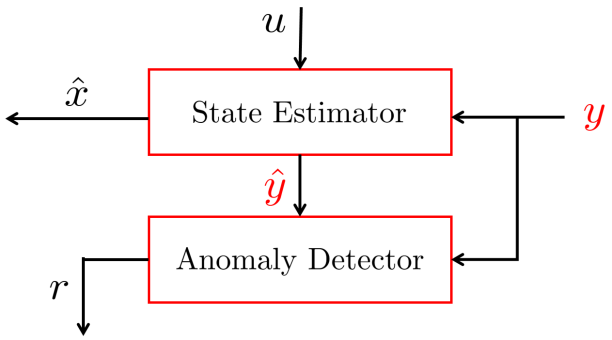


Figure 1.2: Common anomaly detector of control systems

of the sampler and that of the ZOH can be a rational number. Dealing with this general class of the sampled-data system is challenging because the sampling times for the actuation and the measurement are often mismatched, and moreover, the case when the sampling period for the ZOH is larger than the sampling period for the sensor should also be handled. To tackle this issue, we develop the concept of *clustering the time* in the lifted system expression.

1.3 Preliminary: Characterization of detectable and undetectable attacks

Throughout the dissertation, we consider a conventional anomaly detector to classify whether malicious attacks are detectable or not. This conventional anomaly detector (Fig. 1.2) is designed to capture the variation of the residual signal that is defined by

$$r(t) := y(t) - \hat{y}(t)$$

where $y(t)$ is the output measurement of the system and $\hat{y}(t)$ is the output estimation, respectively. Here, any types of estimators are available, such as Luenberger observer, Kalman filter, minimum mean squared error estimator, etc. The anomaly detector identifies whether the attack is injected or not by checking the following inequality is satisfied: $|r(t)| \geq$ threshold. Designing the residual signal has two stages: first, the residual signal $r(t)$ is generated to be insensi-

tive to the disturbance, noise, and uncertainties while sensitive to attack or fault. Then, the threshold is determined to avoid false alarms of the anomaly detector. A few ways to obtain the threshold of the anomaly detector can be found in [AENR88, FD94, SSP03], in which a norm-based evaluation and bounded-real lemma are used to compute threshold. Once the threshold is determined, if $|r(t)| \geq$ threshold, the anomaly detector sounds an alarm that there is something wrong in the system. Hence, whether an attack is detectable or not is defined as follows.

Definition 1.3.1. When a non-zero attack $a(t) \not\equiv 0$ is injected into the system, if $|r(t)| \geq$ threshold, the attack is detectable. On the contrary, when $|r(t)| <$ threshold even if an attack, $a(t) \not\equiv 0$, is injected into the system, the attack is undetectable.

In what follows, it is specified that how the attack on the sensor or actuator affects the residual.

Case 1. Sensor attack

Consider an observable continuous-time linear time-invariant system with an attack on sensor:

$$\begin{aligned} \dot{x}(t) &= Ax(t), \\ y(t) &= Cx(t) + a^{\mathsf{s}}(t), \end{aligned} \tag{1.3.1}$$

where $x \in \mathbb{R}^n$ is the state, $y \in \mathbb{R}^q$ is the output of the system, and $a^{\mathsf{s}} \in \mathbb{R}^q$ is the sensor attack, respectively. In order to get the output estimation, Luenberger type observer is employed as follows:

$$\begin{aligned} \dot{\hat{x}}(t) &= A\hat{x}(t) + L(y(t) - \hat{y}(t)), \\ \hat{y}(t) &= C\hat{x}(t), \end{aligned} \tag{1.3.2}$$

where $\hat{x} \in \mathbb{R}^n$ is the state estimation, $\hat{y} \in \mathbb{R}^q$ is the output estimation, and $L \in \mathbb{R}^{n\times q}$ is a gain matrix chosen $A - LC$ being Hurwitz.

With an error variable $e(t) := x(t) - \hat{x}(t)$, the error dynamics and residual

signal is given by

$$\begin{aligned} \dot{e}(t) &= \dot{x}(t) - \dot{\hat{x}}(t) = (A - LC)e(t) - La^{\mathsf{s}}(t), \\ r(t) &= y(t) - \hat{y}(t) = Cx(t) - C\hat{x}(t) + a^{\mathsf{s}}(t) = Ce(t) + a^{\mathsf{s}}(t). \end{aligned} \tag{1.3.3}$$

Here, we do not consider the disturbance or noise but even in this case the threshold not be set zero. This is because the initial condition of the error dynamics generates non-zero residual value at the transient time. Hence, the threshold should be set to satisfy

$$\text{threshold} > C\|e(0)\|.$$

On the other hand, it is noted that the residual $r(t)$ is directly affected to the attack $a^{\mathsf{s}}(t) \not\equiv 0$, meaning that without sophisticated design of attack $a^{\mathsf{s}}(t)$, most attack signals can be easily detected by the anomaly detector[2].

Case 2. Actuator attack

Let us consider an observable continuous-time linear time-invariant system with an actuator attack:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Ba^{\mathsf{a}}(t), \\ y(t) &= Cx(t), \end{aligned} \tag{1.3.4}$$

where $x \in \mathbb{R}^n$ is the state, $a^{\mathsf{a}} \in \mathbb{R}^p$ is the actuator attack, and $y \in \mathbb{R}^q$ is the compromised output of the system, respectively. Now, as a nominal system, consider the system without the attack.

$$\begin{aligned} \dot{x}_{\mathsf{o}}(t) &= Ax_{\mathsf{o}}(t), \\ y_{\mathsf{o}}(t) &= Cx_{\mathsf{o}}(t), \end{aligned} \tag{1.3.5}$$

where $x_{\mathsf{o}} \in \mathbb{R}^n$ is the state and $y_{\mathsf{o}} \in \mathbb{R}^q$ is the output of the attack-free system. Then, the output of the compromised system (1.3.4) can be rewritten as the summation of the nominal output (1.3.5) and the zero-state output of (1.3.4):

[2]One of the famous sophisticated method satisfying $a^{\mathsf{s}}(t) \not\equiv 0$ but $r(t) \equiv 0$ is the zero dynamics sensor attack introduced in Chapter 3.

that is,

$$y(t) := y_{\mathsf{o}}(t) + y_{\text{att}}(t) = Cx_{\mathsf{o}}(t) + C\int_0^t e^{A(t-\tau)}Ba^{\mathsf{a}}(\tau)d\tau,$$

where $y_{\text{att}}(t) := C\int_0^t e^{A(t-\tau)}Ba^{\mathsf{a}}(\tau)d\tau$.

In view of the defender, the compromised system (1.3.4) can be regarded as

$$\begin{aligned}\dot{x}(t) &= Ax(t),\\ y(t) &= Cx(t) + y_{\text{att}}(t),\end{aligned} \tag{1.3.6}$$

because the defender is unaware of the presence of the actuator attack and the only thing that can be obtained is output measurements. Then, similar to (1.3.3), one has the error dynamics given by

$$\begin{aligned}\dot{e}(t) &= (A - LC)e(t) - Ly_{\text{att}}(t),\\ r(t) &= Ce(t) + y_{\text{att}}(t).\end{aligned} \tag{1.3.7}$$

Similar again, it is noted that almost all non-zero $y_{\text{att}}(t)$ are easily revealed by the anomaly detector since it disturbs the residual $r(t)$ goes to zero.

However, there is a significant distinction between the sensor and actuator attack. In the case of the sensor attack, zero attack, $a^{\mathsf{s}}(t) \equiv 0$, is not effective attack to the system. On the other hand, in the actuator attack case, there is a lethal case even if $y_{\text{att}}(t) \equiv 0$ because the condition $y_{\text{att}}(t) \equiv 0$ does not mean the actuator attack also should be $a^{\mathsf{a}}(t) \equiv 0$. Therefore, all non-zero actuator attacks $a^{\mathsf{a}}(t) \not\equiv 0$ that make the output of the system (1.3.4) become zero (i.e., $y_{\text{att}}(t) \equiv 0$) are undetectable, and also they can be very lethal to the system[3].

Corollary 1.3.1. Non-zero actuator attacks ($a^{\mathsf{a}}(t) \not\equiv 0$) that the effect of them cannot captured in the zero-state output ($y_{\text{att}} <$ threshold) are undetectable attacks. □

[3]The famous attack having this property is the zero dynamics attack, and masking attack in Chapter 4 is also constructed by focusing this point.

Chapter 2

Use of Generalized Hold in Sampled-data Systems to Counteract Zero Dynamics Attack

2.1 Zero Dynamics Attack with Normal Form

The zero dynamics describes the internal behavior of a dynamic system which cannot be observed by the system output, under a particular input that depends on the internal state. The zero dynamics attack arises when this particular input is generated and injected by an attacker, and this attack becomes lethal when the system is non-minimum phase so that the attack signal can drive the system state into an unsafe region of the state-space while such abnormal behavior of the internal state is not detected [TSSJ15, NHV15, PSL+16, PLS18]. In this section, we recall the zero dynamics attack for continuous-time linear systems and for sampled-data systems, and assert that the sampled-data system is more vulnerable to it. We also discuss some existing countermeasures against the zero dynamics attack.

2.1.1 Continuous-time Linear Systems

Consider a SISO linear system given by

$$\begin{aligned} \dot{x}(t) &= Ax(t) + B(u(t) + a^{\mathsf{a}}(t)), \\ y(t) &= Cx(t) \end{aligned} \tag{2.1.1}$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}$ is the control input, $y \in \mathbb{R}$ is the system output, $a^{\mathsf{a}} \in \mathbb{R}$ is the attack signal, and A, B, and C are constant matrices with appropriate dimensions. We can always change the coordinates so that the system (2.1.1) is written as

$$\dot{\eta}(t) = S\eta(t) + P\bar{C}\xi(t), \tag{2.1.2a}$$

$$\dot{\xi}(t) = \bar{A}\xi(t) + \bar{B}\left(\psi^\top \eta(t) + \phi^\top \xi(t) + g(u(t) + a^{\mathsf{a}}(t))\right), \tag{2.1.2b}$$

$$y(t) = \bar{C}\xi(t) \tag{2.1.2c}$$

where $\eta \in \mathbb{R}^{n-\nu}$, $\xi \in \mathbb{R}^{\nu}$, and the parameters S, P, ψ, ϕ, and g are matrices of suitable sizes, and

$$\bar{A} = \begin{bmatrix} 0_{\nu-1} & I_{\nu-1} \\ 0 & 0_{\nu-1}^\top \end{bmatrix}, \ \bar{B} = \begin{bmatrix} 0_{\nu-1} \\ 1 \end{bmatrix}, \ \bar{C} = \begin{bmatrix} 1 & 0_{\nu-1}^\top \end{bmatrix}.$$

The system (2.1.2) is known as *Byrnes-Isidori normal form*; see, e.g., [Kha02] for details. Some important properties of the system can be seen directly from this form. The dimension ν, called *relative degree* of the system, is the number of differentiation of the output until the input u explicitly appears. The sub-dynamics $\dot{\eta}(t) = S\eta(t)$ from (2.1.2) is the *zero dynamics* of the system, which is indeed the internal dynamics when the output y is kept at zero under a particular input. Moreover, it is known that the eigenvalues of S correspond to the zeros of the transfer function of the system (2.1.1). We call the system *minimum phase* if S is Hurwitz, and *non-minimum phase* if S has at least one eigenvalue that has a positive real part. For simplicity, we assume in the sequel that the system has no zero on the imaginary axis.

Suppose that the system (2.1.2) is stabilized by a (dynamic) controller of the form

$$\dot{\zeta}(t) = A_{\mathsf{c}}\zeta(t) + B_{\mathsf{c}}y(t), \tag{2.1.3}$$

$$u(t) = C_{\mathsf{c}}\zeta(t) + D_{\mathsf{c}}y(t) \tag{2.1.4}$$

so that the closed-loop system (2.1.2) and (2.1.3) without attack is exponentially

stable. When the system parameters S, ψ, and g are fully known to the attacker, the zero dynamics attack given by

$$\dot{z}(t) = Sz(t), \tag{2.1.5a}$$

$$a^{\mathsf{a}}(t) = -\frac{1}{g}\psi^\top z(t) \tag{2.1.5b}$$

will steer the state $\eta(t)$ along the trajectory of $z(t)$. Indeed, by subtracting (2.1.5a) from (2.1.2a) with (2.1.5b) plugged into (2.1.2b), one has

$$\begin{aligned}\dot{\eta}(t) - \dot{z}(t) &= S(\eta(t) - z(t)) + P\bar{C}\xi(t),\\ \dot{\xi}(t) &= \bar{A}\xi(t) + \bar{B}\left(\psi^\top(\eta(t) - z(t)) + \phi^\top\xi(t) + gu(t)\right),\\ y(t) &= \bar{C}\xi(t).\end{aligned} \tag{2.1.6}$$

This is a slightly different version of (2.1.2) without attack a^{a} in that the state variable η is changed to $\eta - z$. Hence, the controller (2.1.3) still stabilizes the compromised system (2.1.6), and one can see that

$$\begin{aligned}&\left\|[\eta^\top(t) - z^\top(t), \xi^\top(t), \zeta^\top(t)]^\top\right\|\\ &\qquad\qquad \le \gamma e^{-\lambda t}\left\|[\eta^\top(0) - z^\top(0), \xi^\top(0), \zeta^\top(0)]^\top\right\|\end{aligned} \tag{2.1.7}$$

where γ and λ are positive constants. This implies that, while $\xi(t)$ and $\zeta(t)$ converge to zero, the internal state $\eta(t)$ follows the attacker's state $z(t)$. (For more details, see [PSL+16, Proposition 1].)

Property (2.1.7) illustrates the risk of zero dynamics attack. If the system has at least one unstable zero and the non-zero initial condition $z(0) \neq 0$ belongs to the unstable eigenspace of S, then the attack (2.1.5) can drive the state $\eta(t)$ unbounded, while the other states including the output $y(t) = \bar{C}\xi(t)$ converge to zero. Therefore, if the attack is initiated in steady-state (i.e., when the state $\xi(t)$ and so the output $y(t)$ are zero or almost zero) with a sufficiently small $z(0)$, then this attack is hardly detectable from the output because, even if it perturbs the output a little bit at the time of intrusion, $\xi(t)$ and $y(t)$ converge to zero again. Thus, noting corollary 1.3.1, the attack (2.1.5) remains stealthy in the practical

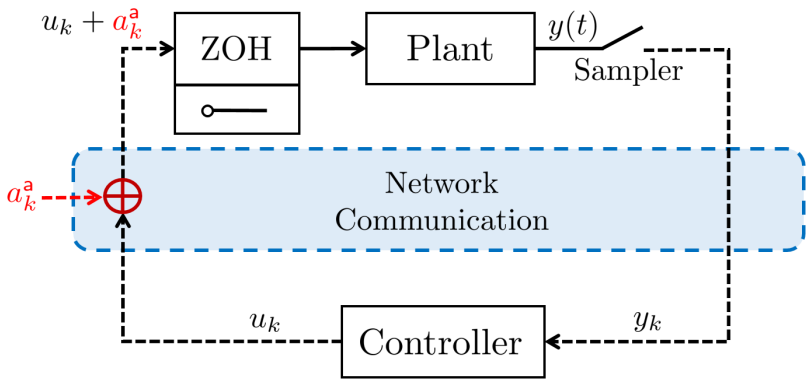


Figure 2.1: Configuration of the compromised sampled-data system with ZOH

sense.

2.1.2 Sampled-data Linear Systems

These days most controllers are implemented digitally while the physical plant is often a continuous-time system. Therefore, the output of the plant is measured at each sampling time kT_s, where k is a nonnegative integer and $T_s > 0$ is the sampling period of the sensor, and the sampled output is utilized to determine the control input that will be used during the next sampling time interval. Under this setting, the sampled output signal is the only available information that can be used to monitor the plant. Correspondingly, it is assumed that the actuator is equipped with a ZOH to generate the continuous-time control input $u(t)$ applied to the continuous-time plant. Specifically, the control signal u_k determined by a controller is sent through a network to the actuator, from which the actual control input is generated by $u(t) := u_k$ for $kT_a \leq t < (k+1)T_a$, where T_a is the sampling period of the ZOH. In general, the sampling time of the sensor and ZOH is the same (i.e., $T_a = T_s$) so that we use T_s as a representative notation throughout this chapter.

Now, we assume that the network is compromised so that the actuator receives a corrupted signal $u_k + a_k^{\mathsf{a}}$ rather than u_k, with a_k^{a} being an attack signal. Hence,

with ZOH, the sampled-data system of (2.1.1) becomes

$$\begin{aligned} x_{k+1} &= A_{\mathsf{d}} x_k + B_{\mathsf{d}}(u_k + a_k^{\mathsf{a}}), \\ y_k &= C_{\mathsf{d}} x_k \end{aligned} \tag{2.1.8}$$

where $x_k = x(kT_s)$, $y_k = y(kT_s)$, u_k, a_k^{a}, $k = 0, 1, \ldots,$ are the state vector, the output, the input and the attack signal, respectively, and

$$A_{\mathsf{d}} := e^{AT_s}, \quad B_{\mathsf{d}} := \int_0^{T_s} e^{A(T_s-\tau)} B d\tau, \quad \text{and} \quad C_{\mathsf{d}} := C.$$

The overall configuration of the system (2.1.8) is depicted in Fig. 2.1.

Similar to the continuous-time system case, the system (2.1.8) can be rewritten as

$$\begin{aligned} \eta_{k+1} &= S_{\mathsf{d}} \eta_k + P_{\mathsf{d}} \bar{C}_{\mathsf{d}} \xi_k \\ \xi_{k+1} &= \bar{A}_{\mathsf{d}} \xi_k + \bar{B}_{\mathsf{d}} \left(\psi_{\mathsf{d}}^\top \eta_k + \phi_{\mathsf{d}}^\top \xi_k + g_{\mathsf{d}}(u_k + a_k^{\mathsf{a}}) \right) \\ y_k &= \bar{C}_{\mathsf{d}} \xi_k \end{aligned} \tag{2.1.9}$$

where $\eta_k \in \mathbb{R}^{n-\mu}$, $\xi_k \in \mathbb{R}^{\mu}$, and the matrices $\bar{A}_{\mathsf{d}}$, $\bar{B}_{\mathsf{d}}$, and $\bar{C}_{\mathsf{d}}$ have the same structures as $\bar{A}$, $\bar{B}$, and $\bar{C}$ with ν being replaced by μ. Note that μ is the relative degree of the sampled-data system (2.1.8) and this may be different from ν even though (2.1.8) is derived from (2.1.1)[1].

As in the continuous-time case, the zero dynamics attack for the sampled-data system (7) is generated by

$$\begin{aligned} z_{k+1} &= S_{\mathsf{d}} z_k, \\ a_k^{\mathsf{a}} &= -\frac{1}{g_{\mathsf{d}}} \psi_{\mathsf{d}}^\top z_k, \end{aligned} \tag{2.1.10}$$

where the initial condition z_0 can be located almost everywhere except the stable eigenspace of S_{d}. If system (2.1.9) has an unstable zero, then the attack (2.1.10), with the initial condition z_0 being chosen such that $\|z_0\|$ is sufficiently small and

[1] Almost all cases, the relative degree of the sampled-data system μ becomes one no matter what the relative degree of the continuous-time system ν was, and more details about this can be found in [YG14, Chapter 3].

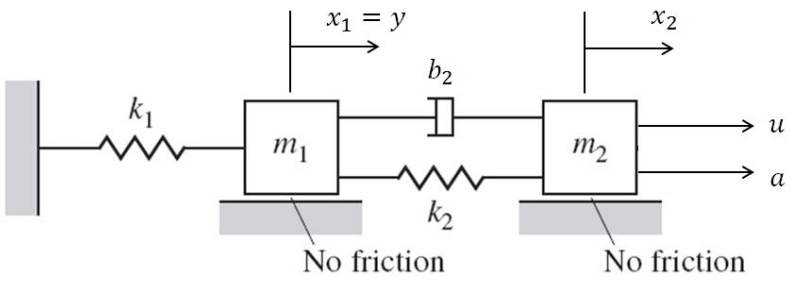


Figure 2.2: Two mass system under actuator attack

exciting the unstable mode of S_{d}, drives the state η_k unbounded while the output y_k is maintained small so that the attack is not detected.

It is emphasized that the sampled-data system with the ZOH can become non-minimum phase even though the corresponding continuous counterpart is minimum phase. In fact, when the system (2.1.1) has relative degree ν greater than two, and the ZOH has sufficiently small sampling period, it is unavoidable that the sampled-data system becomes a non-minimum phase system because of the additionally appearing 'sampling zeros' [sHS84, YG14]. This means that control systems with digital controllers are possibly more vulnerable to the zero dynamics attack than those with continuous-time controllers.

2.1.3 Simulation Result: Zero Dynamics Attack on Sampling Zeros

Consider a two-mass system shown in Fig. 2.2. Two objects with masses m_1 and m_2 are connected through springs (with spring constants k_1 and k_2) and a damper (with damping constant b_2). Assuming that $m_1 = m_2 = 1$kg, $b_2 = 1$Ns/m, and $k_1 = k_2 = 1$N/m, we obtain the transfer function from $u(t)$ to $y(t)$ as

$$G(s) = \frac{(s+1)}{(s^2+1)(s^2+s+1)+(s+1)s^2}$$

Then, $G(s)$ has all poles in the open left half complex plane and one zero at -1. Hence, it is a minimum phase system with relative degree 3. The corresponding sampled-data system with $T_s = 0.1$s can be represented as the normal form (2.1.9)

with

$$S_{\mathsf{d}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.86 & 2.58 & -2.99 \end{bmatrix},\ P_{\mathsf{d}} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},\ \psi_{\mathsf{d}} = \begin{bmatrix} 5.02 \\ 20.04 \\ -28.28 \end{bmatrix},$$
$$\phi_{\mathsf{d}} = 6.78,\ g_{\mathsf{d}} = 1.62 \times 10^{-4}, \tag{2.1.11}$$

and its relative degree is 1. Since the eigenvalues of the matrix S_{d} are -3.64, -0.26, and 0.90, the sampled-data system is not minimum phase anymore. The discrete-time zero dynamics attack (2.1.10), with $z_0 = 10^{-5} \times [1 \quad 0 \quad 0]^\top$, is applied for this system. Fig. 2.3 shows the continuous-time state trajectory and the output signal, and Fig. 2.4 shows the sampled output measurements. Even though the state variables as well as the system output diverge, the sampled output is maintained at almost zero, which shows that the zero dynamics attack can hardly be detected. ◇

2.1.4 Existing Countermeasures Against Zero Dynamics Attack

Since the zero dynamics attack (2.1.10) is hardly detected from the sampled output y_k, it is impossible for classical fault detection schemes, which are usually designed using the measured information of the output, to reveal this stealthy adversary (see section 1.3). With particular attention on the nature of the zero dynamics attack, (only a few) alternative remedies have been introduced in the literature which are briefly discussed in this subsection.

One possible method is to modify the plant's structure by changing actuators and sensors [TSSJ12], or by pre-installing an additional modulation block in front of the control input [HZ16]. In both works, the plant (2.1.8) after the modification is changed into a new dynamics

$$x_{k+1} = \tilde{A}_{\mathsf{d}} x_k + \tilde{B}_{\mathsf{d}}(u_k + a_k^{\mathsf{a}}), \tag{2.1.12}$$
$$y_k = \tilde{C}_{\mathsf{d}} x_k \tag{2.1.13}$$

where the triplet $(\tilde{A}_{\mathsf{d}}, \tilde{B}_{\mathsf{d}}, \tilde{C}_{\mathsf{d}})$ may be unknown to the attacker and is different

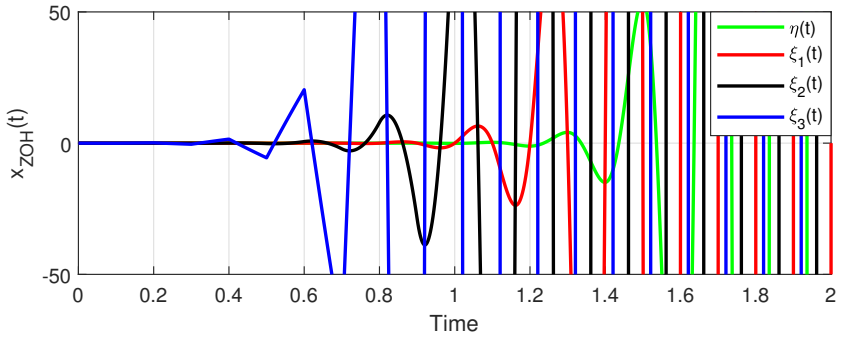


Figure 2.3: Continuous-time state trajectory under sampling zero dynamics attack. $\xi_1(t)$ (red) is the output $y(t)$

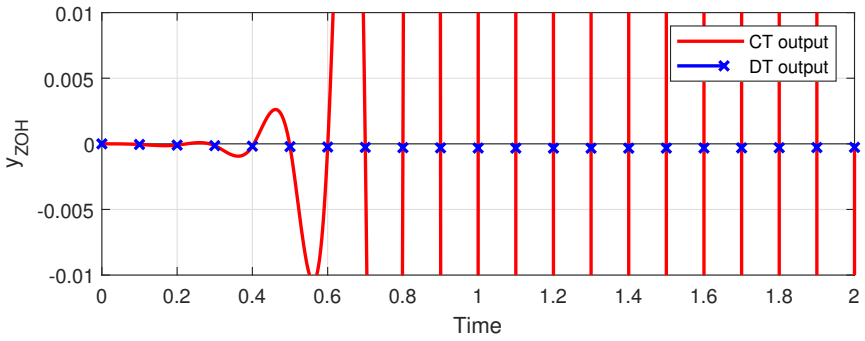


Figure 2.4: Continuous-time output $y(t)$ (red line), which is $\xi_1(t)$ of Fig. 2.3, and sampled output $y(kT_s)$ (blue cross) under the zero dynamics attack.

from the original one $(A_{\mathsf{d}}, B_{\mathsf{d}}, C_{\mathsf{d}})$. In doing so, the zeros of the modified plant (2.1.12) are shifted from their original locations to somewhere unknown to the adversary, and thus the design of the zero dynamics attack (2.1.10) is not possible anymore. However, the modified plant (10) may become vulnerable to the stealthy attacks again whenever the continuous-time plant remains non-minimum phase, or the relative degree is still greater than two and the sampling period is sufficiently small. In fact, it would be possible for the attacker to obtain a rough knowledge of the modified plant (10) by observing its input-to-output response. Then as studied in the recent work [PSL$^+$16], lack of the model knowledge can be overcome by the attackers who obtain the disclosure resources (i.e., measurement output and control input) and use robust control schemes in the design of their attack

strategy (which is the so-called *robust zero dynamics attack*). Moreover, as far as the sampling zeros are concerned, the information on their location can be inferred easily. In fact, the location of sampling zeros depends only on the relative degree (for sufficiently small sampling period T_s) and as $T_s \to 0$ they converge to the roots of Euler-Frobenius polynomial [YG14]. This means that attackers may know the (approximate) location of unstable zeros even after the modification of plant parameters.

Another countermeasure is based on the multi-rate sampler, as introduced by [NHV15]. The underlying idea is to measure the continuous-time output more frequently than the holding action for the discrete-time input, and to construct the stacked measurement

$$\bar{y}_k := \mathrm{col}\big(y(kT_s), y(kT_s + (1/m)T_s), \ldots, y(kT_s + ((m-1)/m)T_s)\big),$$

where m is a positive integer, and regard it as a new output of the system. With sufficiently large m, [NHV15] showed that the multi-rate system from u_k to $\bar{y}_k$ has no discrete-time zero outside the unit circle, no matter where the zeros of the original sampled-data model (2.1.8) are located. Therefore, even if the adversary can redesign the zero dynamics attack to the multi-rate system for stealthiness, its effect becomes less harmful because the resulting attack is generated by a stable dynamics and thus it converges to zero. Although the multi-rate sampler approach of [NHV15] makes it impossible to generate a diverging attack being completely stealthy, this method may not be very effective when the attacker aims to construct a "practically" stealthy attack. A notable attack scenario in this direction would be when the zero dynamics attack targets on the intrinsic zeros (i.e., the discrete-time zeros of (2.1.8) corresponding to the continuous-time zeros of the plant), and the sampling period T_s is small enough. A discrete-time zero dynamics attack exploiting intrinsic zeros can be viewed as an approximated version of its (ideally stealthy) continuous-time counterpart, and the approximation becomes more accurate as the sampling period gets smaller. It means that, with sufficiently small T_s, the difference between the continuous-time outputs under the continuous-time and the discrete-time attacks becomes negligible in a practi-

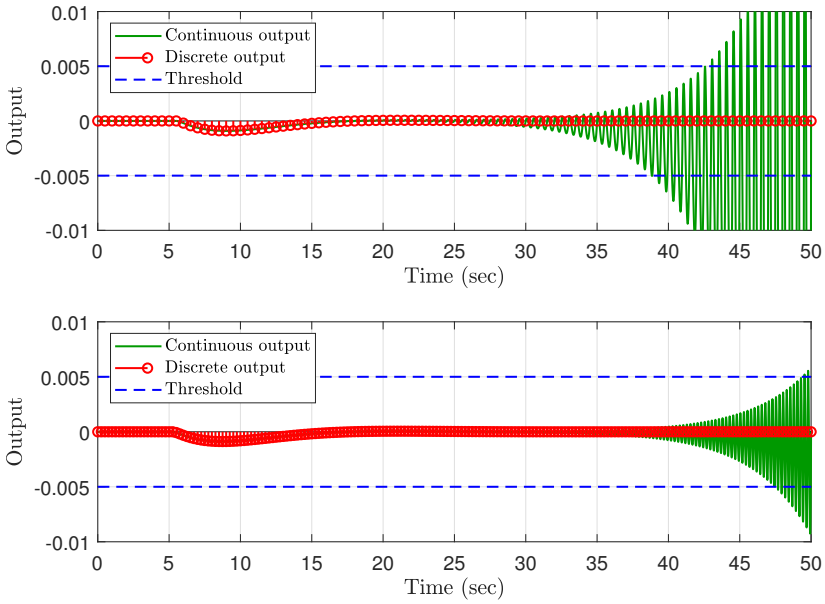


Figure 2.5: Continuous and discrete-time outputs under the zero dynamics attack using intrinsic zeros. The sampling periods are 0.5s (upper) and 0.25s (lower), respectively. Multi-rate sampler will measure the (green) output inbetween the (red) circles.

cal sense. As an illustrative example for this claim, Fig. 2.5 depicts the simulation results with two different sampling periods when a discrete-time zero dynamics attack is made on a hydro turbine [PSL$^+$16]. From this, one can see that the continuous-time output reaches a given threshold more slowly as the sampling period gets smaller; in other words, the discrete-time zero dynamics attack can be (not exactly but) "practically" stealthy for a certain amount of time.

2.2 Optimal Generalized Hold Function to Neutralize Zero Dynamics Attack

As seen from Example 2.1.3, the zero dynamics attack to the sampled-data system becomes effective when the continuous-time system itself contains a non-minimum phase zero, or when a non-minimum phase sampling zero appears. Since

this attack is not detectable, it would be very hard to protect the system as long as the non-minimum phase zeros are present. Thus, if we can move these undesirable zeros of the sampled-data system inside the unit circle in the complex plane, then it is expected that the effect of the zero dynamics attack becomes less harmful, which is the main idea of this paper. To realize the idea, in this section, we employ the 'generalized hold' [Kab87, YG14] instead of the popular ZOH. After showing that we can find a generalized hold to place the zeros at desired locations, we present an optimal design of the hold function considering the inter-sample behavior.

2.2.1 Shifting discrete-time zeros by generalized hold

We first take a look at the basic idea of zero assignment via the generalized hold. Consider a function $h_g(t)$ such that $h_g(t) = 0$ if $t < 0$ or $t \geq T_s$. Also, consider a hold device that generates the continuous-time input to the plant, with a given input sequence u_k, as

$$u(t) = \sum_{k=-\infty}^{\infty} h_g(t - kT_s)u_k.$$

We call this device a *generalized hold with impulse response* $h_g(t)$. Recall that if $h_g(t)$ is a function such that $h_g(t) = 1$, $0 \leq t < T_s$, and $h_g(t) = 0$ otherwise, then the hold device with this impulse response is nothing but the ZOH. With the generalized hold with impulse response $h_g(t)$, the sampled-data model of (2.1.1), free of attack, becomes

$$\begin{aligned} x_{k+1} &= A_{\mathsf{d}}x_k + B_g u_k, \\ y_k &= C_{\mathsf{d}}x_k \end{aligned} \tag{2.2.1}$$

where $x_k \in \mathbb{R}^n$,

$$A_{\mathsf{d}} = e^{AT_s},\ B_g := \int_0^{T_s} e^{A(T_s-\tau)}Bh_g(\tau)d\tau,\ \text{and } C_{\mathsf{d}} = C.$$

The overall block diagram of the system with generalized hold is illustrated in Fig. 2.6. The sampled-data transfer function from u_k to y_k, denoted by $G_{\mathsf{d}}(z)$, is

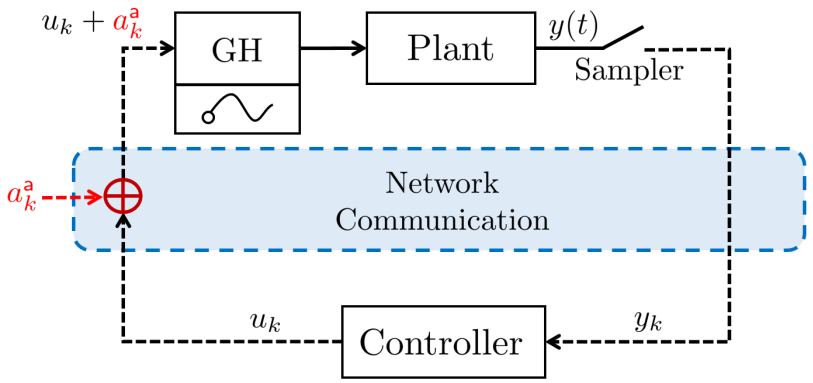


Figure 2.6: Configuration of the compromised sampled-data system with generalized hold

then given by $G_{\mathsf{d}}(z) = C_{\mathsf{d}}(zI_n - A_{\mathsf{d}})^{-1}B_g$.

Now, we investigate the existence of a function $h_g(t)$ which assigns the zeros of the sampled-data transfer function $G_{\mathsf{d}}(z)$ to desired locations. Suppose we would like to make $G_{\mathsf{d}}(z)$ identical to

$$G_{\mathsf{d}}^*(z) = k_{\mathsf{d}} \frac{(z - z_{\mathsf{d},1}) \cdots (z - z_{\mathsf{d},n-1})}{\det(zI_n - A_{\mathsf{d}})} \tag{2.2.2}$$

where $z_{\mathsf{d},1}, \ldots, z_{\mathsf{d},n-1} \in \mathbb{C}$ are the desired zeros (in complex conjugate pairs) and k_{d} is a gain. In what follows, we address the existence of B_g such that $G_{\mathsf{d}}(z)$ becomes $G_{\mathsf{d}}^*(z)$ and then find a function $h_g(t)$ which yields B_g.

Lemma 2.2.1. Suppose $(A_{\mathsf{d}}, C_{\mathsf{d}})$ is observable. Then, there exists $B_g \in \mathbb{R}^n$ such that $G_{\mathsf{d}}(z)$ is identical to $G_{\mathsf{d}}^*(z)$. □

Proof: The proof begins by realizing $G_{\mathsf{d}}^*(z)$ in the controllable canonical form given by

$$\begin{aligned}
\bar{x}_{k+1} &= \left[\begin{array}{c|c} 0_{n-1} & I_{n-1} \\ \hline -d_0 \quad \cdots & -d_{n-1} \end{array}\right] \bar{x}_k + \begin{bmatrix} 0_{n-1} \\ 1 \end{bmatrix} u_k \\
&=: A_{\mathsf{con}} \bar{x}_k + B_{\mathsf{con}} u_k \\
y_k &= \begin{bmatrix} c_0 & \cdots & c_{n-2} & c_{n-1} \end{bmatrix} \bar{x}_k =: C_{\mathsf{con}} \bar{x}_k
\end{aligned}$$

where the constants $c_0, \ldots, c_{n-1}, d_0, \ldots, d_{n-1}$ are determined from the relations

$$\begin{aligned} \det(zI_n - A_{\mathsf{d}}) &= z^n + d_{n-1}z^{n-1} + \cdots + d_0 \\ k_{\mathsf{d}} \prod_{i=1}^{n-1} (z - z_{\mathsf{d},i}) &= c_{n-1}z^{n-1} + c_{n-2}z^{n-2} + \cdots + c_0 \end{aligned} \tag{2.2.3}$$

so that $G^*_{\mathsf{d}}(z) = C_{\mathsf{con}}(zI_n - A_{\mathsf{con}})^{-1}B_{\mathsf{con}}$. Note that $G_{\mathsf{d}}(z)$ is identical to $G^*_{\mathsf{d}}(z)$ if and only if $C_{\mathsf{d}}A^{k-1}_{\mathsf{d}}B_g = C_{\mathsf{con}}A^{k-1}_{\mathsf{con}}B_{\mathsf{con}},\ k = 1, 2, \ldots, n$. This is equivalent to that $\mathcal{O}_{\mathsf{d}}B_g = \mathcal{O}_{\mathsf{con}}B_{\mathsf{con}}$ where $\mathcal{O}_{\mathsf{d}}$ is the observability matrix of $(A_{\mathsf{d}}, C_{\mathsf{d}})$ and $\mathcal{O}_{\mathsf{con}}$ is that of $(A_{\mathsf{con}}, C_{\mathsf{con}})$. Thus, one has

$$B_g = \mathcal{O}^{-1}_{\mathsf{d}}\mathcal{O}_{\mathsf{con}}B_{\mathsf{con}} \tag{2.2.4}$$

which completes the proof. ■

Lemma 2.2.1 states the existence of B_g for zero assignment under the observability condition. With B_g obtained above, it remains to construct $h_g(t)$ such that

$$B_g = \int_0^{T_s} e^{A(T_s-\tau)}Bh_g(\tau)d\tau. \tag{2.2.5}$$

If a function $h_g(t)$ is a solution to (2.2.5), we say $h_g(t)$ is a realization of B_g. It is noted from (2.2.5) that the problem to find $h_g(t)$ can be seen as the problem of finding a control input $h_g(t)$, $0 \le t < T_s$, to steer the state of a system from the origin to B_g at $t = T_s$.

Lemma 2.2.2. Suppose (A, B) of the system (2.1.1) is controllable and $B_g \in \mathbb{R}^n$ is given. Then, there exists a function $h_g(t)$ which is a realization of B_g. □

Proof: With the controllability Gramian

$$W(0, T_s) = \int_0^{T_s} e^{A\tau}BB^\top e^{A^\top\tau}d\tau, \tag{2.2.6}$$

the function

$$h_g(t) = B^\top e^{A^\top(T_s-t)}W^{-1}(0, T_s)B_g \tag{2.2.7}$$

proves the assertion. ■

The generalized hold $h_g(t)$ given in (2.2.7) is a continuous function of t. In

some cases, however, a piecewise constant function h_g is preferred, as in [YG14], in order to implement easily in a digital device. Then, we can also find another realization h_g of B_g, which is piecewise constant and has N subintervals, namely

$$h_g(t) = h_i, \ \frac{(i-1)T_s}{N} \le t < \frac{iT_s}{N}, \quad i = 1, \dots, N. \tag{2.2.8}$$

Substituting (2.2.8) into (2.2.5), one has

$$B_g = \sum_{l=1}^{N} h_l \int_{\frac{(l-1)T_s}{N}}^{\frac{lT_s}{N}} e^{A(T_s-\tau)} B d\tau \tag{2.2.9}$$

which can be written as

$$B_g = \left[A_{\mathsf{d},N}^{N-1} B_{\mathsf{d},N} \quad \cdots \quad A_{\mathsf{d},N} B_{\mathsf{d},N} \quad B_{\mathsf{d},N} \right] h =: \mathcal{C}_{\mathsf{d},N} h \tag{2.2.10}$$

where $h = [h_1, \cdots, h_N]^\top$, $\mathcal{C}_{\mathsf{d},N} \in \mathbb{R}^{n \times N}$ and

$$A_{\mathsf{d},N} = e^{A\frac{T_s}{N}}, \quad B_{\mathsf{d},N} = \int_0^{\frac{T_s}{N}} e^{A(\frac{T_s}{N}-\tau)} B d\tau. \tag{2.2.11}$$

Lemma 2.2.3. For the linear system (2.1.1), consider $A_{\mathsf{d},N}$ and $B_{\mathsf{d},N}$ given by (2.2.11). Suppose $(A_{\mathsf{d},N}, B_{\mathsf{d},N})$ is controllable and $B_g \in \mathbb{R}^n$ is given. Then, the piecewise constant function $h_g(t)$ with $N \ge n$, whose gain h is obtained from $h = \mathcal{C}_{\mathsf{d},N}^{\dagger} B_g$, is a realization of B_g. □

Proof: We first note that the controllability of $(A_{\mathsf{d},N}, B_{\mathsf{d},N})$ ensures that, for any given B_g, there exists a vector h which solves (2.2.10). Since the equation (2.2.10) is just a compact expression of (2.2.9), any solution to (2.2.10) is a solution to (2.2.9), which completes the proof. ■

Remark 2.2.1. If the locations for the desired zeros are determined, then all $z_{\mathsf{d},i}$ in (2.2.2) are decided. But, one still needs to choose k_{d} in (2.2.2) in order to follow the recipe of (2.2.4) and (2.2.7), or (2.2.4), (2.2.8), and (2.2.10). For this, we propose to choose k_{d} such that the integral of $h_g(t)$ for one sampling period is

the same as that of the ZOH, i.e.,

$$\int_0^{T_s} h_g(t)dt = \int_0^{T_s} 1dt = T_s, \text{ or } \sum_{i=1}^{N} h_i = 1_N^\top h = N$$

for the case of (2.2.8) and (2.2.10) where $1_N := [1, \cdots, 1]^\top \in \mathbb{R}^N$. To implement this idea, we propose the following procedure. First, find c_i's such that

$$c_{n-1}z^{n-1} + c_{n-2}z^{n-2} + \cdots + c_1 z + c_0 = \Pi_{i=1}^{n-1}(z - z_{\mathsf{d},i})$$

with $k_{\mathsf{d}} = c_{n-1} = 1$ in (2.2.3). With them, construct C_{con} and $\mathcal{O}_{\mathsf{con}}$ as in the proof of Lemma 2.2.1, which are denoted by $\bar{C}_{\mathsf{con}}$ and $\bar{\mathcal{O}}_{\mathsf{con}}$, respectively. Next, scale the function $h_g(t)$. For example, when $h_g(t)$ is piecewise constant, the gain h of (2.2.8) is taken as

$$h = \frac{NC_{\mathsf{d},N}^\dagger \mathcal{O}_{\mathsf{d}}^{-1} \bar{\mathcal{O}}_{\mathsf{con}} B_{\mathsf{con}}}{1_N^\top C_{\mathsf{d},N}^\dagger \mathcal{O}_{\mathsf{d}}^{-1} \bar{\mathcal{O}}_{\mathsf{con}} B_{\mathsf{con}}}$$

so that $1_N^\top h = N$. □

2.2.2 Design of optimal generalized hold function with security guaranteed

In order to design the generalized hold as discussed in the previous subsection, one needs to choose the desired zeros first and computes corresponding B_g, and then h_g in order. However, in that case, the obtained generalized hold function $h_g(t)$ may happen to be quite different from the gain of the ZOH which is 1, and this causes unintended fluctuation in the inter-sample behavior of the sampled-data system. Thus, recalling our original purpose that is not to assign the zeros at specific locations but to make the system have stable zeros, we propose to formulate an optimization problem to minimize the unnecessary fluctuation under the constraint that the zeros are located inside the unit circle.

In this regard, we formulate a performance index consisting of the difference of state trajectories with the ZOH and the generalized hold; that is, $x_{\mathrm{gh}}(t) - x_{\mathrm{zoh}}(t)$.

The sampled step responses (i.e., when $u_k \equiv 1$) of $x_{\text{zoh}}(t)$ and $x_{\text{gh}}(t)$ are given by

$$x_{\text{zoh}}(t) = e^{At}x_{\text{zoh}}(0) + \int_0^t e^{A(t-\tau)}Bd\tau, \tag{2.2.12}$$

$$x_{\text{gh}}(t) = e^{At}x_{\text{gh}}(0) + \int_0^t e^{A(t-\tau)}B\sum_{k=0}^{\infty} h_g(\tau - kT_s)d\tau. \tag{2.2.13}$$

Under the same initial conditions, the difference between (2.2.12) and (2.2.13) follows as

$$x_{\text{gh}}(t) - x_{\text{zoh}}(t) = \int_0^t e^{A(t-\tau)}B\big(\sum_{k=0}^{\infty} h_g(\tau - kT_s) - 1\big)d\tau. \tag{2.2.14}$$

However, as seen in (2.2.14), it is quite tricky to use the difference of state trajectories as it is. So, as a bypass, we deal with this minimization problem in the frequency domain, not in the time domain. Next, we consider the continuous-time frequency response of sampled data systems, and for this, we recall the result of [FG94]. First, rewrite the generalized hold $h_g(t)$ as a periodic function; namely,

$$\bar{h}_g(t) = \sum_{k=-\infty}^{\infty} h_g(\tau - kT_s),$$

which can be represented as a Fourier series of the form

$$\bar{h}_g(t) = \sum_{p=-\infty}^{\infty} a_p^{\text{gh}} e^{jpw_0 t}$$

where $w_0 = 2\pi/T_s$. Then, the input matrix with the generalized hold, $B_g = \int_0^{T_s} e^{A(T_s-\tau)}Bh_g(\tau)d\tau$, can be rewritten using the coefficients of the Fourier series as follows[2]:

$$B_g = \sum_{p=-\infty}^{\infty} a_p^{\text{gh}} B_p \tag{2.2.15}$$

where

$$B_p = \int_0^{T_s} e^{(A-jpw_0 I)\tau}d\tau B.$$

[2]The detailed procedure can be found in [FG94, Lemma 3.1]

Before we go further, we define the frequency response of the system as

$$X(w) = (jwI_n - A)^{-1}BU(w) =: H(w)U(w)$$

and let $U_s(w)$ denoting the transform of the sampled input

$$u_s(t) := \sum_{k=-\infty}^{\infty} \delta(t - kT_s)u_k.$$

Now, the continuous-time frequency response of the sampled-data system with generalized hold can be computed by

$$X_{\text{gh}}(w) = \frac{1}{2\pi}H(w)\{\bar{H}_g(w) * [H_{\text{zoh}}(w)U_s(w)]\} \tag{2.2.16}$$

where $H_{\text{zoh}}(w) = \frac{1-e^{-jwT_s}}{jw}$ and $\bar{H}_g(w)$ denotes the Fourier transform of $\bar{h}_g(t)$, i.e., $\bar{H}_g(w) = 2\pi\sum_{p=-\infty}^{\infty} a_p^{\text{gh}}\delta(w - pw_0)$. Substituting (2.2.15) in (2.2.16), one has[3]

$$X_{\text{gh}}(w) = \sum_{p=-\infty}^{\infty} a_p^{\text{gh}} H_{\text{zoh}}(w - pw_0)H(w)U_{\text{s}}(w)$$

and when the case with the zero-order hold is given by

$$X_{\text{zoh}}(w) = a_0^{\text{zoh}} H_{\text{zoh}}(w)H(w)U_{\text{s}}(w)$$

where $a_0^{\text{zoh}} = a_0^{\text{gh}}$. In turn, the difference of the frequency response between the generalized hold and zero-order hold is given by

$$X_{\text{gh}}(w) - X_{\text{zoh}}(w) = \left\{ \sum_{p=-\infty}^{\infty} a_p^{\text{gh}} H_{\text{zoh}}(w - pw_0) - a_0^{\text{gh}} H_{\text{zoh}}(w) \right\} H(w)U_{\text{s}}(w). \tag{2.2.17}$$

Taking a close look the equation (2.2.17) (or (2.2.15)), the whole information associated with the generalized hold is contained in the coefficients a_p^{gh}, and it is clear that as the coefficients $a_p^{\text{gh}}, p \neq 0$ get smaller, $\|X_{\text{gh}}(w) - X_{\text{zoh}}(w)\|$ becomes

[3]The details can be found in [FG94, Theorem 3.1]

small. Hence, in the problem to find a generalized hold function $h_g(t)$ minimizing $\|X_{\rm gh}(w) - X_{\rm zoh}(w)\|$, the performance index can be switched into the following one:

$$\left\| \sum_{p=-\infty}^{\infty} a_p^{\rm gh} B_p - a_0^{\rm gh} B_0 \right\| = \|B_g - B_{\mathsf{d}}\| = \left\| \int_0^{T_s} e^{A(T_s-\tau)} B(h_g(\tau) - 1) d\tau \right\|. \tag{2.2.18}$$

Of course, this is not the equivalent to $\|X_{\rm gh}(w) - X_{\rm zoh}(w)\|$ but as long as the high frequency component of $H(w)$ is not dominant, we can use (2.2.18) in the sense that the generalized hold that minimizes (2.2.18) is expected to minimize $\|X_{\rm gh}(w) - X_{\rm zoh}(w)\|$ because both of them depend on how large $a_p^{\rm gh}, p \neq 0$ is.

On the other hand, in order to allocate the discrete-time zeros inside the unit circle, we employ one of the following two lemmas as a constraint of the optimization problem:

Lemma 2.2.4. [[Mor84, Theorem 1]] The polynomial

$$c_{n-1} z^{n-1} + c_{n-2} z^{n-2} + \cdots + c_0$$

is Schur stable if

$$\sum_{i=0}^{n-2} |c_i| < |c_{n-1}|.$$

□

Lemma 2.2.5. [Jury's stability test] A second order polynomial

$$E(z) := z^2 + c_1 z + c_0$$

is Schur stable if and only if the following conditions hold:

$$E(1) > 0, \quad E(-1) > 0, \quad |c_0| < 1.$$

□

Jury's stability test is necessary and sufficient condition for Schur stability but it does not provide a convex set when the dimension of the zero polynomial

is larger than 2. So in that case, Mori's lemma can be used as a substitute (throughout the rest of this section, we use the latter for consistency). Next, as a second constraint, the following constraint is imposed to keep the DC gain of the ZOH equal to that of the generalized hold, which is given by

$$\int_0^{T_s} h_g(t)dt = T_s. \tag{2.2.19}$$

Now, we consider the case where the generalized hold gain h_g is parameterized by a finite dimensional vector. In particular, we take the coefficients of the numerator of the desired sampled-data transfer function $G_{\mathsf{d}}^*(z)$ as the parameter

$$\bar{c} := [c_0, \cdots, c_{n-1}]^\top = C_{\mathsf{con}}^\top$$

(see (2.2.3)). Then, the matrix B_g can be expressed as a linear function of $\bar{c}$ as follows:

$$\begin{aligned} B_g &= \mathcal{O}_{\mathsf{d}}^{-1}\mathcal{O}_{\mathsf{con}}B_{\mathsf{con}} = \mathcal{O}_{\mathsf{d}}^{-1}\begin{bmatrix} C_{\mathsf{con}}B_{\mathsf{con}} \\ \vdots \\ C_{\mathsf{con}}A_{\mathsf{con}}^{n-1}B_{\mathsf{con}} \end{bmatrix} \\ &= \mathcal{O}_{\mathsf{d}}^{-1}\begin{bmatrix} B_{\mathsf{con}}^\top C_{\mathsf{con}}^\top \\ \vdots \\ B_{\mathsf{con}}^\top (A_{\mathsf{con}}^\top)^{n-1}C_{\mathsf{con}}^\top \end{bmatrix} = \mathcal{O}_{\mathsf{d}}^{-1}\begin{bmatrix} B_{\mathsf{con}}^\top \\ \vdots \\ B_{\mathsf{con}}^\top (A_{\mathsf{con}}^\top)^{n-1} \end{bmatrix}\bar{c} =: S\bar{c} \end{aligned} \tag{2.2.20}$$

in which, we used the fact that $C_{\mathsf{con}}A_{\mathsf{con}}^{i-1}B_{\mathsf{con}}$ is scalar since the system has single input and single output. Hence, the performance index (2.2.18) becomes

$$\|B_g - B_{\mathsf{d}}\| = \|S\bar{c} - B_{\mathsf{d}}\|, \tag{2.2.21}$$

which is convex function respect to $\bar{c}$. Subsequently, the second constraint (2.2.19) also can be rewritten as a convex constraint by two ways. First, using (2.2.20) with (2.2.7) in Lemma 2.2.2, it follows that

$$h_g(t) = B^\top e^{A^\top(T_s-t)}W^{-1}(0,T_s)S\bar{c} =: \mathcal{S}_1(t)\bar{c},$$

by which (2.2.19) rewritten as

$$\left(\int_0^{T_s} \mathcal{S}_1(t)dt\right)\bar{c} = T_s. \tag{2.2.22}$$

The other one is to use Lemma 2.2.3. From h_i in (2.2.8), one has

$$h = \mathcal{C}_{\mathsf{d},N}^{\dagger} B_g = \mathcal{C}_{\mathsf{d},N}^{\dagger} S\bar{c} =: \mathcal{S}_2\bar{c}$$

or $h_i = \mathcal{S}_{2,i}\bar{c}$ where $\mathcal{S}_{2,i}$ is the i-th row of the matrix $\mathcal{S}_2$. This implies

$$\left(\sum_{i=1}^{N} \mathcal{S}_{2,i}\right)\bar{c} = N. \tag{2.2.23}$$

At last, with the cost function (2.2.21), Lemma 2.2.4, and (2.2.23) (or (2.2.22)) the optimization problem is formulated as

$$\underset{\bar{c}=(c_0,\cdots,c_{n-1})}{\text{minimize}} \ \|S\bar{c} - B_{\mathsf{d}}\| \tag{2.2.24a}$$

$$\text{subject to } \sum_{i=0}^{n-2} |c_i| \le |c_{n-1}| - \epsilon, \tag{2.2.24b}$$

$$\text{and } \left(\sum_{i=1}^{N} \mathcal{S}_{2,i}\right)\bar{c} = N \text{ (or (2.2.22))} \tag{2.2.24c}$$

in which, small positive number ϵ is introduced to implement the strict inequality of Lemma 2.2.4.

Remark 2.2.2. Although the constraints (2.2.24b) make the problem (2.2.24) not convex, there is a simple workaround. Since the constraint can be divided into two convex constraints:

$$\sum_{i=0}^{n-2} |c_i| \le c_{n-1} - \epsilon \quad \text{with} \quad c_{n-1} > 0, \tag{2.2.25}$$

$$\sum_{i=0}^{n-2} |c_i| \le -c_{n-1} - \epsilon \quad \text{with} \quad c_{n-1} < 0, \tag{2.2.26}$$

one can solve two sets of optimization problems, each of which has (2.2.25) or

(2.2.26) instead of (2.2.24b) for (2.2.24). Then, compare two performance indices, and take the minimal one. Using a convex optimization solver (e.g., `cvxgen`), the optimal value of each problem can be easily obtained. □

Remark 2.2.3. As a matter of fact, the condition on Lemma 2.2.4 may seem quite conservative. On the other hand, "Jury's stability test" provides a necessary and sufficient condition for Schur stability of a polynomial. Hence, it would be the most suitable constraint for guaranteeing the Schur stability in the proposed optimization problem. Unfortunately, however, the set of coefficients from Jury's test is not convex in general[4], which makes the optimization problem (2.2.24) non-convex. To go around this non-convexity problem, we employ a convex-type sufficient (but not too conservative) condition for Schur stability, that is, the condition in Lemma 2.2.4. With that as a constraint, the proposed optimization problem (2.2.24) has been cast into a convex optimization problem, which can be easily solved by `cvxgen`.

To verify how conservative the condition of Lemma 2.2.4 is, we compare the condition used in Lemma 2.2.4 with that of Jury's test by showing the sets of coefficients satisfying those conditions. To show the sets graphically, we consider a simple case with a monic polynomial of degree 3 given by $\mathcal{P}(z) = z^3 + a_2 z^2 + a_1 z + a_0$. The conditions for $\mathcal{P}(z)$ being Schur stable are given as follows:

C-1 *Jury's test*:
(i) $\mathcal{P}(1) > 0$, (ii) $\mathcal{P}(-1) < 0$, (iii) $|a_0| < 1$, (iv) $|a_0^2 - 1| > |a_0 a_2 - a_1|$

C-2 *Condition in Lemma 2.2.4*: $\sum_{i=0}^{2} |a_i| < 1$

The coefficients tuple $(a_0, a_1, a_2) \in [-5, 5]^3$ satisfying C-1 and C-2 are plotted in Fig. 2.7. It is noted that the set obtained from C-1 (blue in Fig. 2.7) is obviously non-convex, whereas that from C-2 (red in Fig. 2.7) is convex. On the other hand, the sets in Fig. 2.7 are merged, and its projections to a_0–a_1, a_0–a_2, and a_1–a_2 planes, respectively, are illustrated in Fig. 2.8 to show how conservative C-2 is with respect to C-1. □

[4]When the degree of the polynomial is 2, the set of coefficients from Jury's test is a convex set.

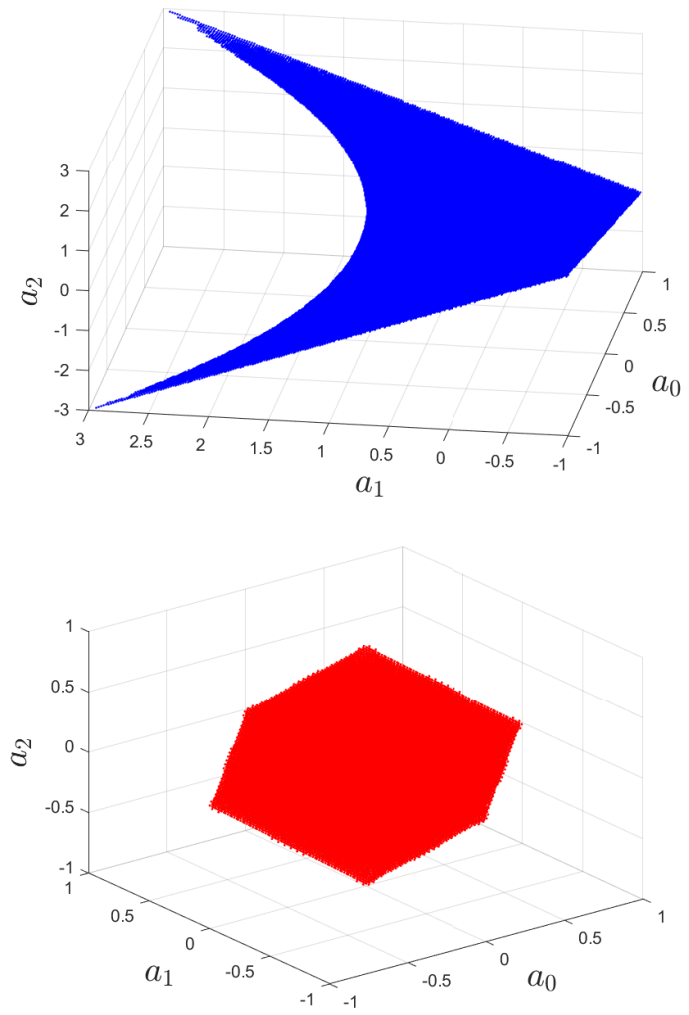


Figure 2.7: Sets of the coefficient tuple (a_0, a_1, a_2) for C-1 (upper) and C-2 (lower)

2.2.3 Simulation Results: Effect of Optimal Generalized Hold

Although the system considered in Example 2.1.3 is minimum phase in the continuous-time domain, the sampled-data system under the ZOH became a non-minimum phase system due to the sampling zero at -3.64. We would like to move the zeros inside the unit circle in the complex plane for neutralizing the zero dy-

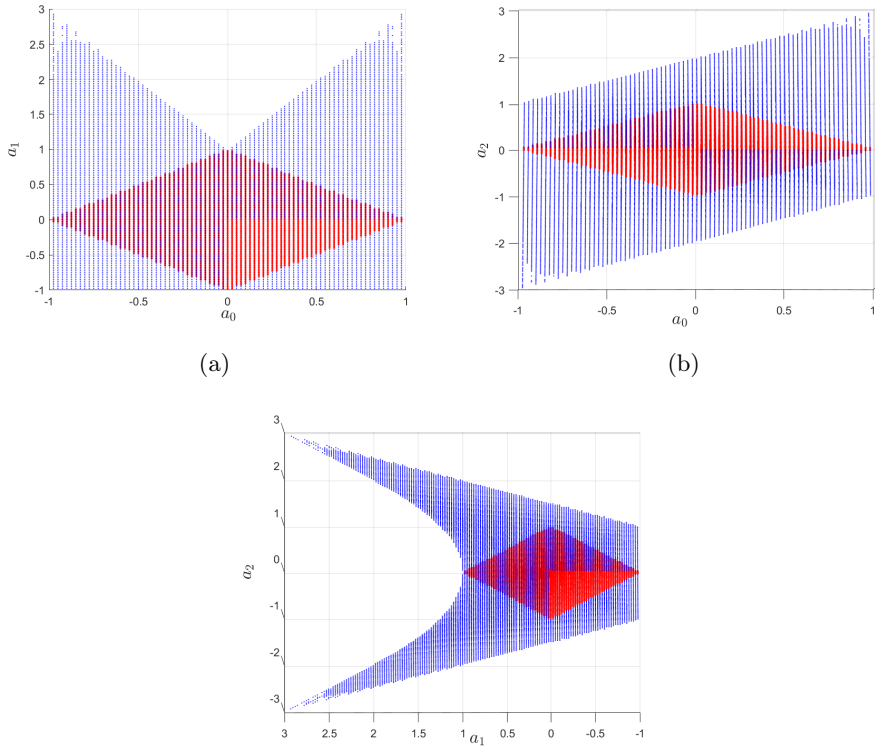


Figure 2.8: Projections to (a) a_0–a_1, (b) a_0–a_2, and (c) a_1–a_2 planes of the merged set

namics attack by employing the proposed optimal piecewise constant generalized hold with $N = 4$ subintervals (since $n = 4$). Setting $\epsilon = 10^{-5}$, the optimal generalized hold is given as

$$h = [5.86, \ -0.70, \ -2.31, \ 1.15]^\top, \tag{2.2.27}$$

by which the discrete-time zeros are shifted to

$$z_{\mathsf{d},1} \approx -0.99, \quad z_{\mathsf{d},2} \approx e^{-T_s}, \quad \text{and} \quad z_{\mathsf{d},3} \approx 0.$$

With the generalized hold having hold gain (2.2.27), we have the sampled-

data system whose normal form is given as (2.1.9) with

$$S_{\mathsf{d}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0.90 & -0.09 \end{bmatrix},\ P_{\mathsf{d}} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},\ \psi_{\mathsf{d}} = \begin{bmatrix} -0.82 \\ 6.95 \\ -6.68 \end{bmatrix},$$
$$\phi_{\mathsf{d}} = 3.89,\ g_{\mathsf{d}} = 4.75 \times 10^{-4},\ \mu = 1. \tag{2.2.28}$$

Now, it is expected that the previous zero dynamics attack based on ZOH model (2.1.11) is not stealthy anymore. On the other hand, even if the new shifted zeros (or, the model (2.2.28)) are revealed to the attackers, their zero dynamics attack is not effective anymore because *new zeros are all stable*. In fact, Fig. 2.9 demonstrates that all states converge to zero when the sampling zero dynamics attack (2.1.10), which uses the model (2.2.28) with $z_0 = 10^{-5} \times [1 \quad 0 \quad 0]^\top$, is injected into the system. Fig. 2.10 shows the inter-sample behaviors of the system when the unit step input is injected. Specifically, at the top of Fig. 2.10, in order to assign zeros at specific location ($z_{\mathsf{d},1} = e^{-T_s},\ z_{\mathsf{d},2} = 0,\ z_{\mathsf{d},3} = 0$), the generalized hold designed by Lemma 2.2.3 is used, with which h^* is obtained as

$$h^* = [20.88,\ -21.97,\ 3.14,\ 1.94]^\top$$

(it is obtained without using the optimization). On the other hand, the proposed optimal generalized hold (**??**) is used in the middle of Fig. 2.10, and the ZOH is employed in the bottom of that figure. In comparison to the case of ZOH (bottom), one can see that the inter-sample behavior of the optimal generalized hold (middle) is more desirable than that of the non-optimal one (top).

2.3 Illustrative Example for Closed-loop System

In this section, we consider a worktable motion control system that appears in [DB08, Example 13.10]. The transfer function of the continuous-time plant is described by

$$G(s) = \frac{1}{s(s+10)(s+20)},$$

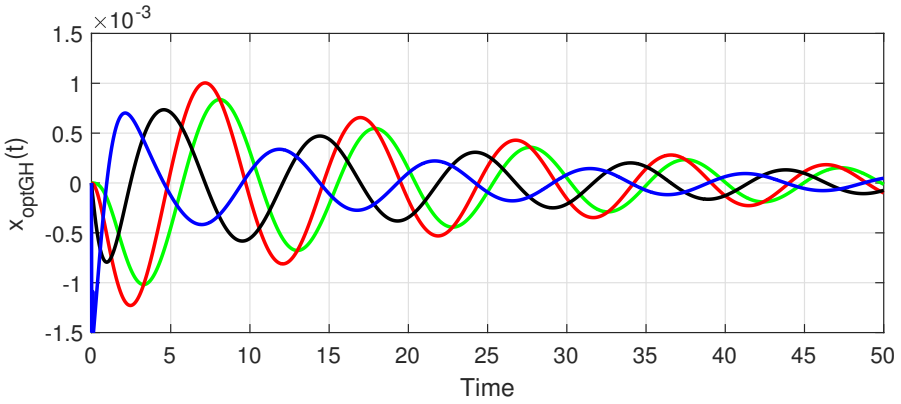


Figure 2.9: Continuous-time state trajectory under the zero dynamics attack with shifted zeros. Green line is the output $y(t)$. The sampled output is not drawn due to the large time scale.

and a lead compensator

$$K(s) = 8000\frac{(s+11)}{(s+62)}$$

is proposed to meet the design specifications such as a rise time of 0.25s with an overshoot less than 5%. With $T_s = 1/100$s, the ZOH equivalent model $G_{\mathsf{d}}(z)$ of $G(s)$ is represented by the normal form (2.1.9) with

$$S_{\mathsf{d}} = \begin{bmatrix} 0 & 1 \\ -0.86 & -3.71 \end{bmatrix},\ P_{\mathsf{d}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},\ \psi_{\mathsf{d}} = \begin{bmatrix} -4.8 \\ -25.51 \end{bmatrix},$$
$$\phi_{\mathsf{d}} = 6.44,\ g_{\mathsf{d}} = 1.55 \times 10^{-7},\ \mu = 1, \tag{2.3.1}$$

and $K(s)$ is discretized via the pole-zero matching method as

$$K(z) = 6293\frac{(z - e^{-11T_s})}{(z - e^{-62T_s})}.$$

The initial states of the system are given as $\eta_0 = (0, 0)$ and $\xi_0 = 0.2$. Note that the relative degree of the sampled-data system $G_{\mathsf{d}}(z)$ is 1, while that of the continuous-time plant $G(s)$ is 3. Thus, its sampled-data system has two sampling zeros, and their locations are -0.25 and -3.47, respectively. As we expected, the sampling zeros include unstable one (that is -3.47) so that the system (2.3.1) is

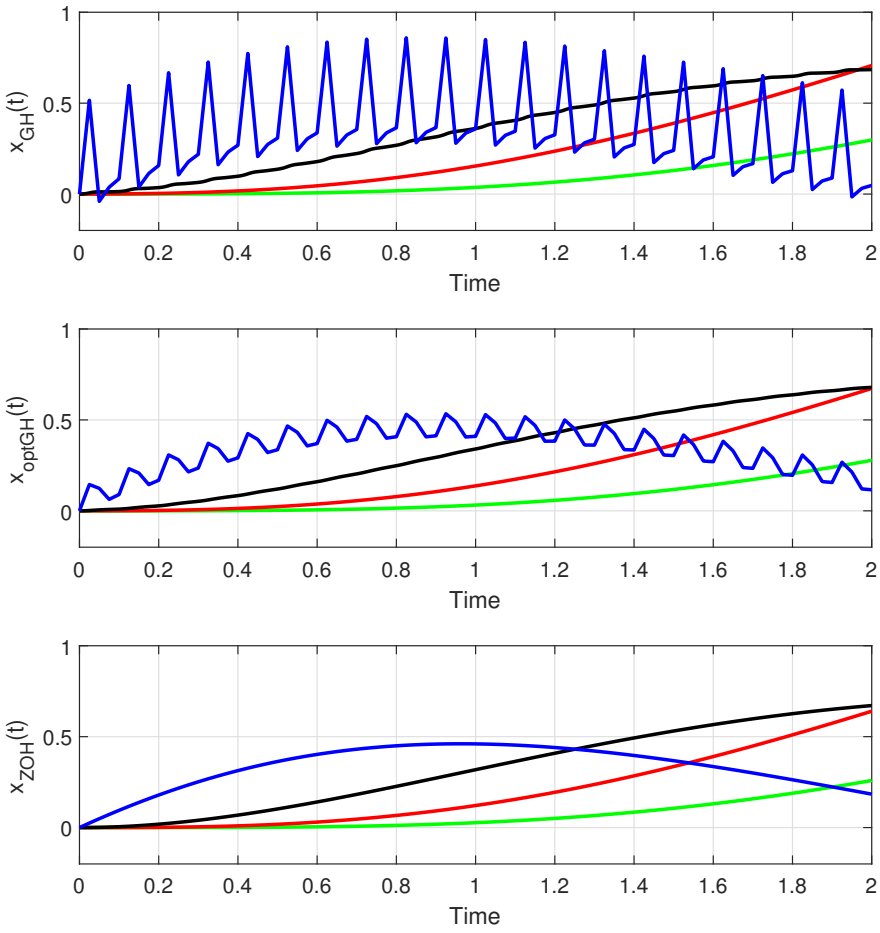


Figure 2.10: Continuous-time state trajectory under the unit step input without attack when the generalized hold with desired locations of zeros, $z_{\mathsf{d},1} = e^{-T_s}$, $z_{\mathsf{d},2} = 0$, $z_{\mathsf{d},3} = 0$ (top), the optimal GH (middle), and ZOH are employed (bottom), respectively.

vulnerable to the sampling zero dynamics attack.

Now, obtaining the optimal generalized hold from (**??**), we neutralize the zero dynamics attack. The obtained optimal generalized hold is given by $h = [4.99, -2.72, 0.72]^\top$, by which the discrete-time zeros are shifted to $z_{\mathsf{d},1} \approx -0.99$ and $z_{\mathsf{d},2} \approx 0$, respectively. Then, the resultant discrete-time system is now mini-

mum phase, whose normal form (2.1.9) becomes

$$S_{\mathsf{d}} = \begin{bmatrix} 0 & 1 \\ 0 & -0.99 \end{bmatrix},\ P_{\mathsf{d}} = \begin{bmatrix} 0 \\ 1 \end{bmatrix},\ \psi_{\mathsf{d}} = \begin{bmatrix} 0.74 \\ -6.19 \end{bmatrix}, \tag{2.3.2}$$
$$\phi_{\mathsf{d}} = 3.72,\ g_{\mathsf{d}} = 4.31 \times 10^{-7},\ \mu = 1,$$

and the initial states are the same as those of (2.3.1). Usually, the digital controller $K(z)$ may have to be reconfigured according to the change of the holding device. In this example, however, we used the same digital controller $K(z)$ since it still guarantees the desired design specifications.

We now investigate the safety of the discrete-time system (2.3.1) and (2.3.2) against the zero dynamics attack (2.1.10). In Fig. 2.11, the upper one shows the output of the discrete-time system (2.3.1) under the zero dynamics attack (2.1.10). As we expected before, the attack can not be detected from the sampled output because the sampled output remains near zero while the continuous output is diverging. Meanwhile, to generate the zero dynamics attack for the system (2.3.2), the attackers have no choice but to use S_{d} of (2.3.2), which is stable now, and thus, the attack signal can be interpreted as nothing but a vanishing perturbation. Therefore, we can easily guess that the output of the system will converge to zero and this is actually demonstrated in the lower part of Fig. 2.11 where both the continuous and sampled outputs converge to zero.

2.4 Experiment: DC/DC Converter with Electrical Circuit

In this section, we apply the proposed strategy to a DC/DC converter with electric circuit. Fig. 2.12 shows the overall configuration of the experimental equipment, which is composed of a DC/DC converter, electrical circuit, sensor, and micro control unit (MCU). Specifically, the equipment works as follows: to begin with, the DC/DC converter converts 48V DC voltage (exogenously supplied power) into the desired DC current with the range of $0 \sim 30$A, and the specific value of the desired current is determined corresponding to the voltage signal

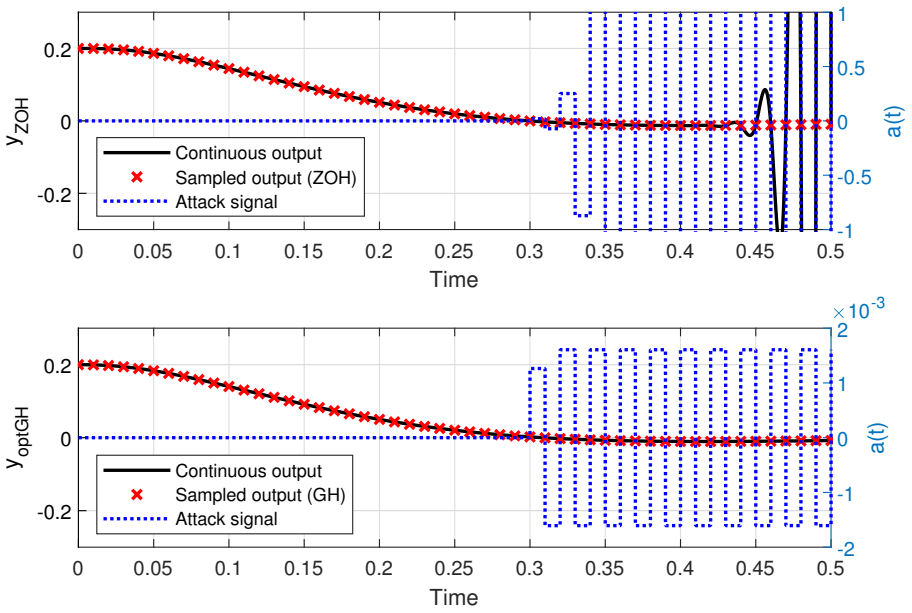


Figure 2.11: Output of the worktable motion control system under zero dynamics attack when the ZOH (upper) and optimal generalized hold (lower) are used, respectively. The attack is injected into the system at 0.3sec.

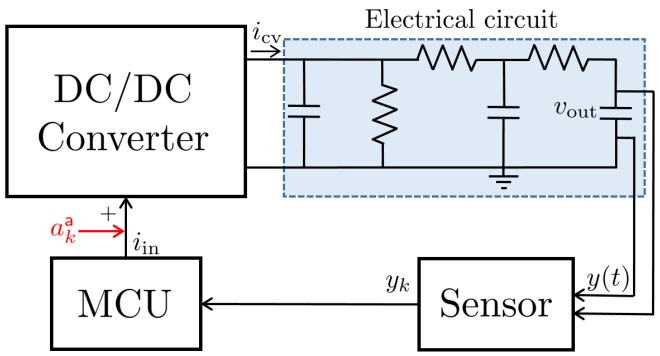


Figure 2.12: Overall configuration of the DC/DC converter with electrical circuit

$0 \sim 3$V of the MCU operating with pulse width modulation (PWM) signal. For instance, when the PWM is set as 50%, the output voltage of the MCU becomes

1.5V, which drives the output of the converter as 7.5A. Meanwhile, the electrical circuit is designed as a high order low-pass filter (LPF) whose input and output are the output current of the converter and a load voltage of a capacitor, respectively. Lastly, the sensor captures the voltage of the capacitor and transmits them to the MCU every 0.001sec (i.e., $T_s = 0.001$sec or 1kHz).

In what follows, the specific system model of the converter and electrical circuit are given. In order to match the converted current $i_{\rm cv}$ with the command signal of the MCU, we use the current range $0 \sim 30$A (in the command line) as the control signal of the MCU instead of the voltage range $0 \sim 3$V. Then, the transfer function of the converter from $i_{\rm in}$ to $i_{\rm cv}$ (see. Fig. 2.12) is given by

$$G_{\rm cv}(s) = \frac{2.28 \times 10^{-8} s + 0.0024}{1.3 \times 10^{-24} s^4 + 2.3 \times 10^{-18} s^3 + 7.6 \times 10^{-13} s^2 + 9.5 \times 10^{-8} s + 0.0024}, \tag{2.4.1}$$

and the transfer function of the electrical circuit (from $i_{\rm cv}$ to $v_{\rm out}$ in Fig. 2.12) is given as follows:

$$G_{\rm ec}(s) = \frac{1}{7.39 \times 10^{-10} s^3 + 5.22 \times 10^{-6} s^2 + 8.74 \times 10^{-3} s + 2}. \tag{2.4.2}$$

By combining (2.4.1) and (2.4.2), one has the overall transfer function of the system whose input and output are $i_{\rm in}$ and $v_{\rm out}$, respectively; that is,

$$\begin{aligned} G_{\rm ov}(s) &= G_{\rm cv}(s) G_{\rm ec}(s) \\ &= \frac{2.28 \times 10^{-8} s + 0.0024}{\mathsf{D}_7 s^7 + \mathsf{D}_6 s^6 + \mathsf{D}_5 s^5 + \mathsf{D}_4 s^4 + \mathsf{D}_3 s^3 + \mathsf{D}_2 s^2 + \mathsf{D}_1 s^1 + \mathsf{D}_0}, \end{aligned} \tag{2.4.3}$$

where

$$\begin{aligned} &\mathsf{D}_7 = 9.9 \times 10^{-34},\ \mathsf{D}_6 = 1.7 \times 10^{-27},\ \mathsf{D}_5 = 5.7 \times 10^{-22},\ \mathsf{D}_4 = 7 \times 10^{-17} \\ &\mathsf{D}_3 = 2.3 \times 10^{-12},\ \mathsf{D}_2 = 1.3 \times 10^{-8},\ \mathsf{D}_1 = 2.1 \times 10^{-5},\ \mathsf{D}_0 = 0.0048. \end{aligned}$$

Hence, the relative degree ν of the system (2.4.3) is $\nu = 6$, and the poles and

zeros of the system (2.4.3) are given as:

$$\begin{aligned}&\text{Zeros: } -1.05\times 10^5\\ &\text{Poles: } -1.32\times 10^6,\ (-0.18+0.1i)\times 10^6,\ (-0.18-0.1i)\times 10^6,\\ &\qquad -0.03\times 10^6,\ -0.05\times 10^5,\ -0.02\times 10^5,\ -0.03\times 10^4.\end{aligned} \tag{2.4.4}$$

We note that even though the system (2.4.3) is stable minimum phase system, since the relative degree ν is larger than 2, with a sufficiently small sampling times, the associated sampled-data system becomes non-minimum phase so that it is vulnerable to the sampling zero dynamics attack.

For more efficient experiment, when we design the zero dynamics attack and generalized hold, we will use the approximated system having a lower dimension compared to the original one (2.4.3). This is possible because the system is composed of extremely fast poles and zeros as seen in (2.4.4). In particular, the poles $-1.32\times 10^6,\ (-0.18+0.1i)\times 10^6,\ (-0.18-0.1i)\times 10^6$ are much faster compared to the others, which means the deviation between the original system (2.4.3) and the approximated one without consideration of such fast poles is trivial. Likewise, the zero is also very fast and stable so that its effect on the step response is restricted. In this context, an approximated system can be considered, in which the zero and fastest 3 poles in G_{ov} are removed as follows:

$$G_{\text{app}}(s)=\frac{1}{2.22\times 10^{-14}s^4+8.96\times 10^{-10}s^3+5.48\times 10^{-6}s^2+8.8\times 10^{-3}s^1+2} \tag{2.4.5}$$

whose poles are $-0.03\times 10^6,\ -0.05\times 10^5,\ -0.02\times 10^5,\ -0.03\times 10^4$.

Remark 2.4.1. It is one thing to note that the relative degree of the original plant (2.4.3) is $\nu=6$ but that of the altered one is now 4, denoted by $\bar{\nu}=4$. Thus, when we design the sampling zero dynamics attack, one might have a concern that the location of the sampling zeros when $T_s\to 0$ depends on the relative degree of the associated continuous-time plant so that the constructed sampling zero dynamics attack based on $G_{\text{app}}(s)$ may not be stealthy. However, this is the case when the sampling time T_s is sufficiently small. In the experiment settings, the sampling

time is set up as $T_s = 0.001$sec, and with this setting, the location of the sampling zeros of the sample-data system more relies on the continuous-time poles and zeros than the relative degree. Indeed, the sampling zeros of $G_{\mathrm{ov}}(s)$ (2.4.3) are given by $-1.029,\ -0.038,\ -1.91\times10^{-5},\ 1.21\times10^{-17},\ -1.04\times10^{-17},\ -7.6\times10^{-21}$, while that of $G_{\mathrm{app}}(s)$ are given by $-1.029,\ -0.039,\ -2.0\times10^{-5}$. Thus, the adversaries can generate the sampling zero dynamics attack with the approximated system $G_{\mathrm{app}}(s)$. □

2.4.1 Simulation Results

We now carry out a computer simulation with MATLAB/Simulink to verify the dangerousness of the sampling zero dynamics attack and the effectiveness of the proposed generalized hold. In what follows, as discussed in the previous section, we generate the zero dynamics attack based on the approximated system $G_{\mathrm{app}}(s)$, and then it is injected into the original system $G_{\mathrm{ov}}(s)$ to show its lethality. Subsequently, we design the proposed optimal generalized hold to counteract the zero dynamics attack. To this end, with the ZOH and sensor whose sampling period is 0.001sec, the associated sampled-data model of $G_{\mathrm{app}}(s)$ can be computed as follows:

$$
\begin{aligned}
&S_{\mathsf{d}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -0.0 & -0.04 & -1.07 \end{bmatrix},\ P_{\mathsf{d}} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},\ \psi_{\mathsf{d}} = \begin{bmatrix} -0.0 \\ -0.08 \\ -2.15 \end{bmatrix}, \\
&\phi_{\mathsf{d}} = 1.96, \quad g_{\mathsf{d}} = 1, \quad \mu = 1.
\end{aligned} \tag{2.4.6}
$$

With the information of the sampled-data system (2.4.6) (zeros: $-1.029, -0.039, -2.0\times10^{-5}$), the sampling zero dynamics attack can be generated as

$$
\begin{aligned}
z_{k+1} &= \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -0.0 & -0.04 & -1.07 \end{bmatrix} z_k, \\
a_k^{\mathsf{a}} &= \begin{bmatrix} 0.0 & 0.08 & 2.15 \end{bmatrix} z_k, \quad z_0 = 10^{-4} \times \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}^{\top}.
\end{aligned} \tag{2.4.7}
$$

Now, we set up several simulation specifications. Above all, as we know the

sampling rate of the digital devices (MCU, sensor) is fixed to 1kHz, and in order to observe the influence of the zero dynamics attack effectively, set 15A as the reference input since it is bounded to $0 \sim 30$A corresponding to the spec of the converter. Then, it is initiated to inject the constructed attack into the system (2.4.3) through the MCU when $t = 0.01$sec (this is waiting time for capacitor charging). The simulation results using MATLAB/Simulink are illustrated in Fig. 2.13. It is clear that the generated attack is disruptive (diverging) as seen in Fig. 2.13-(a) since the zero dynamics S_{d} has unstable zero at -1.029. As one can see in Fig. 2.13-(b), even though the attack signal a_k^{a} is constructed using the information of the approximated system (2.4.5), the effect of the attack is still undetectable and destructive. However, due to the input saturation, the effect of the attack cannot grow indefinitely, and it reaches the boundary in about 0.35sec.

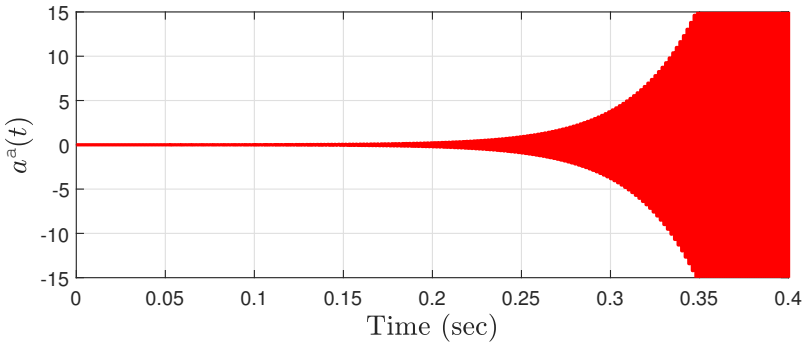
Now, in order to neutralize the zero dynamics attack, we design a piecewise constant optimal generalized hold by following the Section **??**. The resultant generalized hold is give by

$$h = [1.005, \quad 1.0033, \quad 0.2152, \quad 1.7809]^\top, \tag{2.4.8}$$

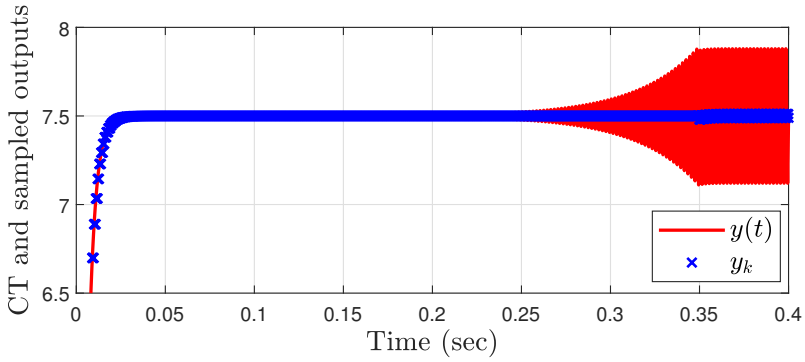
by which the discrete-time zeros of the system are shifted to $z_{\mathsf{d},1} = -0.67$, $z_{\mathsf{d},2} = -0.02$, $z_{\mathsf{d},3} = -0.0$ so that the system becomes minimum phase. Fig. 2.14 shows the result of the use of optimal generalized hold when the zero dynamics attack (2.4.7) is injected. In this case, the exploited information to construct the zero dynamics attack (2.4.7) does not guarantee the stealthiness of the attack so that its effect is captured at the sampled output as one can see in Fig. 2.14.

2.4.2 Experiment Results

The actual experimental setup is pictured in Fig. 2.15. The optimal generalized hold (2.4.8) and zero dynamics attack (2.4.7) is implemented using MCU (TMS320F28335 manufactured by Texas Instruments Incorporated). As a matter of fact, the MCU (DSP chip) conducts the role of both D/A and A/D converter, and the sampling period of the MCU is set as 250μs since the number of the



(a)



(b)

Figure 2.13: Simulation results for: (a) Zero dynamics attack $a^{a}(t)$. (b) Continuous-time output $y(t)$ and sampled output y_k when the zero dynamics attack (2.4.7) is injected into the system.

subintervals of the designed generalized hold is 4 and the the prior sampling period is 0.001s. Under these settings, we performed an experiment in which the proposed strategy is implemented in MCU likewise the MATLAB simulation. The performed experiment results are shown in Fig. 2.16.

We note that there occurred several differences compared to computer simulation. First of all, there is a deviation in the steady-state value of the output; in the computer simulation, the steady-state value is 7.5V but that of the experiment is about 6.8V. This difference is seen as coming from the modeling error. On the other hand, even though Fig. 2.16-(a) shows similar tendency to the computer simulation (see Fig. 2.13-(b)), a few exceptional sampled measurements are

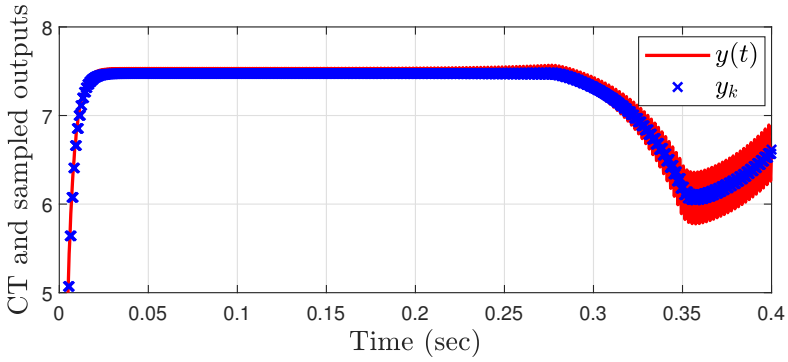


Figure 2.14: Continuous-time output $y(t)$ and sampled output y_k when the zero dynamics attack (2.4.7) is injected into the system equipped with the optimal generalized hold (2.4.8).

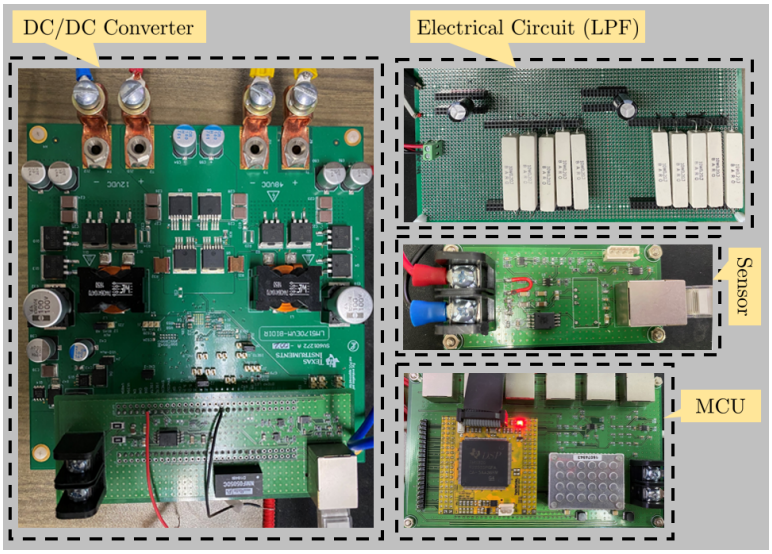


Figure 2.15: Experiment equipment

captured due to the noise of the sensor. Continually, as we saw in the computer simulation, in the result of the experiment Fig. 2.16-(b), we again confirmed that the use of the optimal generalized hold is effective to capture the impact of the zero dynamics attack. Lastly, Fig. 2.16-(c) shows well the harmless of the zero dynamics attack that is constructed using the altered (stable) zeros.

2.5 Study on the Effect of Generalized Hold on Intrinsic Zeros of Nonlinear Systems under Fast Sampling

The threat of the zero dynamics attack does not disappear even we consider a nonlinear system. Indeed, in [PLS18], the authors showed that the uncertain non-minimum phase nonlinear systems are also exposed to a threat of the zero dynamics attack using a quadruple-tank case study. It is noted that the effectiveness of the zero dynamics attack on nonlinear systems is still originated from the non-minimum phaseness of the system as same as linear systems. This means if the zero dynamics of the nonlinear system is stabilized by using the generalized hold function, it is expected to neutralize the zero dynamics attack.

Therefore, in this subsection, we would like to expand our discussions to nonlinear systems in a somewhat restrictive perspective. Specifically, we narrow down our focus to the non-minimum phase sampled-data system whose non-minimum phaseness comes from the continuous-time plant, which means the zero dynamics of the continuous-time nonlinear system is unstable. Furthermore, in order to consider this problem more simply, we suppose that the relative degree of the continuous-time plant is 1 and the sampling period of the digital devices is sufficiently small (fast sampling). We also restrict the form of the generalized hold as a piecewise constant function with 2 subintervals. This is because, when we want to shift intrinsic zeros of the system, it is known to require a huge cost as the sampling time T_s getting smaller (infinitely large hold gain is required) [HS19]. However, in real world, the sampling time T_s is not zero even though it may become very small, so it is realizable theoretically but fluctuations of the inter-sample behavior, the chronic problem of the use of generalized hold, becomes a remarkable issue again. Thus, we use a generalized hold in a restrictive way for attenuating such fluctuations, but in order to modify the zero dynamics part of sampled-data system, the number of subintervals should be at least larger than the relative degree so we use a generalized hold having 2 subintervals. Of course, because of the limitation of the generalized hold, the zero dynamics of the sampled-data system cannot be modified freely. In this context, we investigate how the generalized hold affects the zero dynamics of the system in the state-space

perspective and how much the generalized hold can change the zero dynamics of the sampled-data system.

We start with Byrnes-Isidori normal form as follows:

$$\begin{aligned}\dot{\xi}(t) &= \begin{bmatrix} 0_{\nu-1} & I_{\nu-1} \\ 0 & 0_{\nu-1}^{\top} \end{bmatrix} \xi(t) + \begin{bmatrix} 0_{\nu-1} \\ 1 \end{bmatrix} (\phi(\xi,\eta) + \psi(\xi,\eta)u(t)), \\ \dot{\eta}(t) &= C(\xi,\eta), \\ y(t) &= \xi_1(t),\end{aligned} \tag{2.5.1}$$

where $\xi = \mathrm{col}(\xi_1, \cdots, \xi_\nu) \in \mathbb{R}^\nu$ and $\eta \in \mathbb{R}^{n-\nu}$. We narrow down our focus to a restricted system whose relative degree is 1 (i.e., $\nu = 1$) and the case of input function $\psi(\xi,\eta)$ is a constant ψ, due to the complexity of general nonlinear systems to come. It is also supposed that the system (2.5.1) is non-minimum phase; that is, the zero dynamics $\dot{\eta}(t) = C(0,\eta(t))$ is not stable. Then, with the generalized hold $h_g(t)$, the corresponding system (2.5.1) for each sampling period, $kT_s \le t < (k+1)T_s$, becomes

$$\begin{aligned}\dot{\xi}_1(t) &= \phi(\xi_1,\eta) + \psi(\xi_1,\eta)h_g(t-kT_s)u_k, \\ \dot{\eta}(t) &= C(\xi_1,\eta).\end{aligned} \tag{2.5.2}$$

As aforementioned, we assume the generalized hold $h_g(t)$ has a form of a piecewise constant function whose gains are denoted by h_1 and h_2, respectively, so it follows that

$$h_g(t) := \begin{cases} h_1, & 0 \le t < \frac{T_s}{2}, \\ h_2, & \frac{T_s}{2} \le t < T_s, \\ 0, & \text{otherwise} \end{cases}$$

In addition, it is assumed that $h_1 + h_2 = 2$ as the same reason for the linear case.

In what follows, we compute the sampled-data system of (2.5.2) using *truncated Taylor series* (TTS) method [YG14] with the generalized hold. As a matter of fact, the sampled-data system obtained by using a truncated Taylor series is not the exact discrete system but it is an approximated one because of the *truncation* of the Taylor series expansion. In particular, we truncate Taylor series at the

first order of the sampling time T_s since we consider the fast sampling so that the larger order than 1 can be ignored. By using the TTS method, the approximated sampled-data system (TTS model) is computed as follows:

1. The first sub-interval: $kT_s \le t < kT_s + \frac{T_s}{2}$

$$\begin{aligned}\xi_1(kT_s + \frac{T_s}{2}) &= \xi_1(kT_s) + \frac{T_s}{2}\Big(\phi(\xi_1(kT_s), \eta(kT_s)) + \psi h_1 u_k\Big),\\ \eta(kT_s + \frac{T_s}{2}) &= \eta(kT_s) + \frac{T_s}{2} C(\xi_1(kT_s), \eta(kT_s)).\end{aligned} \tag{2.5.3}$$

2. The second sub-interval: $kT_s + \frac{T_s}{2} \le t < kT_s + T_s$

$$\begin{aligned}\xi_1(kT_s + T_s) &= \xi_1(kT_s + \frac{T_s}{2})\\ &\quad + \frac{T_s}{2}\Big(\phi(\xi_1(kT_s + \frac{T_s}{2}), \eta(kT_s + \frac{T_s}{2})) + \psi h_2 u_k\Big),\\ \eta(kT_s + T_s) &= \eta(kT_s + \frac{T_s}{2}) + \frac{T_s}{2} C(\xi_1(kT_s + \frac{T_s}{2}), \eta(kT_s + \frac{T_s}{2})).\end{aligned} \tag{2.5.4}$$

Now, combining (2.5.3) and (2.5.4), we derive a sampled-data system for the whole sampling period, $kT_s \le t < kT_s + T_s$. To this end, we rewrite functions

$$\phi(\xi_1(kT_s + \frac{T_s}{2}), \eta(kT_s + \frac{T_s}{2})) \quad \text{and} \quad C(\xi_1(kT_s + \frac{T_s}{2}), \eta(kT_s + \frac{T_s}{2}))$$

to become functions of kT_s using truncated Taylor series with the first order again; that is,

$$\begin{aligned}\phi(\xi_1(kT_s + \frac{T_s}{2}), \eta(kT_s + \frac{T_s}{2})) &= \phi(\xi_1(kT_s), \eta(kT_s)) + \frac{T_s}{2}\dot{\phi}(\xi_1(kT_s), \eta(kT_s))\\ &= \phi(\xi_1(kT_s), \eta(kT_s)) + \frac{T_s}{2}\frac{\partial \phi}{\partial z}(kT_s)\begin{bmatrix}\dot{\xi}_1(kT_s)\\ \dot{\eta}(kT_s)\end{bmatrix}\\ C(\xi_1(kT_s + \frac{T_s}{2}), \eta(kT_s + \frac{T_s}{2})) &= C(\xi_1(kT_s), \eta(kT_s)) + \frac{T_s}{2}\frac{\partial C}{\partial z}(kT_s)\begin{bmatrix}\dot{\xi}_1(kT_s)\\ \dot{\eta}(kT_s)\end{bmatrix}\end{aligned} \tag{2.5.5}$$

where $z := [\xi_1 \quad \eta^\top]^\top$. Then, by substituting (2.5.3) and (2.5.5) into (2.5.4), one

has

$$\begin{aligned}
&\xi_1(kT_s+T_s)\\
&\quad=\xi_1(kT_s)+\frac{T_s}{2}\Big(\phi(\xi_1(kT_s),\eta(kT_s))+\psi h_1u_k\Big)\\
&\qquad+\frac{T_s}{2}\left\{\phi(\xi_1(kT_s),\eta(kT_s))+\frac{T_s}{2}\left(\frac{\partial\phi}{\partial z}\right)^{\top}\begin{bmatrix}\dot\xi_1(kT_s)\\ \dot\eta(kT_s)\end{bmatrix}+\psi h_2u_k\right\}\\
&\quad=\xi_1(kT_s)+T_s\phi(\xi_1(kT_s),\eta(kT_s))+\frac{T_s}{2}\psi h_1u_k\\
&\qquad+\frac{T_s^2}{2^2}\frac{\partial\phi}{\partial\xi_1}\Big(\phi(\xi_1(kT_s),\eta(kT_s))+\psi h_1u_k\Big)\\
&\qquad+\frac{T_s^2}{2^2}\left(\frac{\partial\phi}{\partial\eta}\right)^{\top}C(\xi_1(kT_s),\eta(kT_s))+\frac{T_s}{2}\psi h_2u_k,
\end{aligned}\tag{2.5.6}$$

$$\begin{aligned}
\eta(kT_s+T_s)&=\eta(kT_s)+\frac{T_s}{2}C(\xi_1(kT_s),\eta(kT_s))\\
&\qquad+\frac{T_s}{2}\left(C(\xi_1(kT_s),\eta(kT_s))+\frac{T_s}{2}\frac{\partial C}{\partial z}\begin{bmatrix}\dot\xi_1(kT_s)\\ \dot\eta(kT_s)\end{bmatrix}\right)\\
&=\eta(kT_s)+T_sC(\xi_1(kT_s),\eta(kT_s))\\
&\qquad+\frac{T_s^2}{2^2}\left(\frac{\partial C}{\partial\xi_1}\right)\Big(\phi(\xi_1(kT_s),\eta(kT_s))+\psi h_1u_k\Big)\\
&\qquad+\frac{T_s^2}{2^2}\left(\frac{\partial C}{\partial\eta}\right)C(\xi_1(kT_s),\eta(kT_s)).
\end{aligned}\tag{2.5.7}$$

Now, in order to capture the fast dynamics under fast sampling, we transform the system (2.5.6) and (2.5.7) with the delta operator;

$$\delta x:=\frac{x(kT_s+T_s)-x(T_s)}{T_s}.$$

Then, it follows that

$$\begin{aligned}
\delta\xi_1(kT_s)&=\phi(\xi_1(kT_s),\eta(kT_s))+\frac{1}{2}\psi h_1u_k+\frac{T_s}{2^2}\frac{\partial\phi}{\partial\xi_1}\Big(\phi(\xi_1(kT_s),\eta(kT_s))+\psi h_1u_k\Big)\\
&\qquad+\frac{T_s}{2^2}\left(\frac{\partial\phi}{\partial\eta}\right)^{\top}C(\xi_1(kT_s),\eta(kT_s))+\frac{1}{2}\psi h_2u_k,
\end{aligned}\tag{2.5.8}$$

$$\delta\eta(kT_s) = C(\xi_1(kT_s),\eta(kT_s)) + \frac{T_s}{2^2}\left(\frac{\partial C}{\partial \xi_1}\right)\Big(\phi(\xi_1(kT_s),\eta(kT_s)) + \psi h_1 u_k\Big) + \frac{T_s}{2^2}\left(\frac{\partial C}{\partial \eta}\right)C(\xi_1(kT_s),\eta(kT_s)). \tag{2.5.9}$$

It is noted that the equations (2.5.8) and (2.5.9) recover the continuous-time system (2.5.2) when the sampling time $T_s \to 0$. However, our interest is to modify the zero dynamics of the system using the generalized hold function $h_g(t)$ even when $T_s \to 0$, and to accomplish this purpose, it is needed to guarantee that

$$\lim_{T_s\to 0} T_s h_g(t) \neq 0.$$

This means we should use the hold gains of the generalized hold function to become a function of $1/T_s$ such that

$$h_1 := \frac{\bar{h}_1}{T_s} \quad \text{and} \quad h_2 := \frac{\bar{h}_2}{T_s}, \tag{2.5.10}$$

where $\bar{h}_1$ and $\bar{h}_2$ are new design parameters. When appropriate hold gains $\bar{h}_1$ and $\bar{h}_2$ are decided later, the actual generalized hold is determined as follows:

$$h_g(t) = \begin{cases} \frac{\bar{h}_1}{T_s}, & \text{for } \ 0 \le t < \frac{T_s}{2}, \\ \frac{\bar{h}_2}{T_s} = 2 - \frac{\bar{h}_1}{T_s}, & \text{for } \ \frac{T_s}{2} \le t < T_s, \\ 0, & \text{otherwise.} \end{cases}$$

With the generalized hold (2.5.10), the sampled-data system (2.5.8) and (2.5.9) under fast sampling (i.e., the system (2.5.8) and (2.5.9) when $T_s \to 0$) becomes

$$\begin{aligned} \delta\xi_1(kT_s) &= \phi(\xi_1(kT_s),\eta(kT_s)) + \psi u_k + \frac{1}{2^2}\frac{\partial \phi}{\partial \xi_1}\psi \bar{h}_1 u_k, \\ \delta\eta(kT_s) &= C(\xi_1(kT_s),\eta(kT_s)) + \frac{1}{2^2}\frac{\partial C}{\partial \xi_1}\psi \bar{h}_1 u_k. \end{aligned} \tag{2.5.11}$$

In order to find a zero dynamics of (2.5.11), we compute u_k deriving the output

becoming zero (i.e., $y = \xi_1 = 0$), which is obtained by finding u_k satisfying

$$\delta\xi_1 = \phi(\xi_1(kT_s), \eta(kT_s)) + \left\{\psi + \frac{1}{2^2}\frac{\partial\phi}{\partial\xi_1}\psi\bar{h}_1\right\}u_k = 0.$$

So, the input u_k is given by

$$u_k = -\frac{\phi(\xi_1(kT_s), \eta(kT_s))}{\psi + \frac{1}{2^2}\frac{\partial\phi}{\partial\xi_1}\psi\bar{h}_1}. \tag{2.5.12}$$

Then, by substituting (2.5.12) and $\xi_1 = 0$ into (2.5.11), we can obtain the zero dynamics of the sampled-data system with the generalized hold under fast sampling, which is computed by

$$\begin{aligned}
\delta\eta(kT_s) &= C(0, \eta(kT_s)) + \frac{1}{2^2}\left(\frac{\partial C}{\partial\xi_1}\right)\Big|_{\xi_1=0,\eta=\eta(kT_s)}\left(\psi\bar{h}_1\right)u_k \\
&= C(0, \eta(kT_s)) - \frac{\psi\bar{h}_1\phi(\xi_1(kT_s), \eta(kT_s))}{2^2\psi + \frac{\partial\phi}{\partial\xi_1}\psi\bar{h}_1}\frac{\partial C}{\partial\xi_1}\Big|_{\xi_1=0,\eta=\eta(kT_s)} \\
&= C(0, \eta(kT_s)) - \frac{\psi\bar{h}_1}{4\psi + \frac{\partial\phi}{\partial\xi_1}(0, \eta(kT_s))\psi\bar{h}_1}\phi(0, \eta(kT_s))\frac{\partial C}{\partial\xi_1}(0, \eta(kT_s)).
\end{aligned} \tag{2.5.13}$$

Noting that when we determine $\bar{h}_1$, we should not use $\bar{h}_1$ such that

$$4\psi + \frac{\partial\phi}{\partial\xi_1}(0, \eta(kT_s))\psi\bar{h}_1 = 0$$

for avoiding infinitely diverging state. At last, by observing the zero dynamics (2.5.13), one can see that how the hold gain of the generalized hold $\bar{h}_1$ can affect to the zero dynamics (intrinsic zeros) of the sampled-data system.

However, even though we can modify the zero dynamics (2.5.13) by choosing hold gain $\bar{h}_1$, obviously there exists a limit on variation to change the zero dynamics coming from the system structure. In other words, no matter how the gain $\bar{h}_1$ is designed, because of ϕ and C, it may not possible to make the zero dynamics (2.5.13) being stable. This implies that the system class that can be modified into the minimum phase by using a piecewise constant generalized hold is quite restrictive. In order to formalize such a system class, we define the following.

Definition 2.5.1. For the nonlinear system (2.5.2) which is non-minimum phase, if the associated sampled-data system with the generalized hold (2.5.11) can be modified into minimum phase (i.e., the system (2.5.13) becomes asymptotically stable) by designing a generalized hold function $h_g(t)$, then the system (2.5.2) is called '*minimum phasable*'. □

Unfortunately, there is no theoretical condition to classify whether the given system is minimum phasable or not yet. Thus, when a specific non-minimum phase system model is given, one should check manually that the zero dynamics with the generalized hold (2.5.13) can be changed into the minimum phase or not by adjusting $\bar{h}_1$. Although it is quite tricky to check the given system is minimum phasable or not, if it is, the system can be safe from the nonlinear zero dynamics attack.

Remark 2.5.1. When it comes to considering linear systems, the discussions so far can be much simplified and it can be found a sufficient condition for minimum phasable systems. Consider a non-minimum phase linear system with the relative degree 1 written as a normal form as follows:

$$\begin{aligned} \dot{\xi}_1 &= A_c\xi_1 + B_c(\lambda\xi_1 - c_m C_0\eta + c_m u), \\ \dot{\eta} &= A_0\eta + B_0 C_c\xi_1, \\ y &= C_c\xi_1 \end{aligned} \tag{2.5.14}$$

where $A_c = 0, B_c = 1$, and $C_c = 1$. By following the same procedure that is discussed above (see (2.5.3) – (2.5.11)), one has a sampled-data system with generalized hold under fast sampling as follows:

$$\begin{bmatrix} \delta\xi_1(kT_s) \\ \delta\eta(kT_s) \end{bmatrix} = \begin{bmatrix} \lambda & -c_m C_0 \\ B_0 & A_0 \end{bmatrix} \begin{bmatrix} \xi_1(kT_s) \\ \eta(kT_s) \end{bmatrix} + \begin{bmatrix} c_m + \frac{1}{4}\lambda c_m \bar{h}_1 \\ \frac{c_m}{4} B_0 \bar{h}_1 \end{bmatrix} u_k \tag{2.5.15}$$

where $\bar{h}_1 := T_s h_1$. Subsequently, by computing u_k satisfying $\delta\xi_1(kT_s) = 0$ and substituting such u_k into (2.5.15) with $\xi_1 = 0$, the zero dynamics can be obtained as follows.

$$\delta\eta(kT_s) = (A_0 + \frac{\bar{h}_1 c_m}{\bar{h}_1\lambda + 4} B_0 C_0)\eta(kT_s)$$

$$=: (A_0 + \gamma B_0 C_0)\eta(kT_s), \tag{2.5.16}$$

which has exactly the same structure compared to (2.5.13). Moreover, it is noted that the dynamics (2.5.16) can be considered as a closed-loop system with output feedback of the system (A_0, B_0, C_0). This observation provides us a sufficient condition for minimum phasable system of linear systems; that is, if the system (A_0, B_0, C_0) is output feedback stabilizable, then the system (2.5.14) is minimum phasable. □

Remark 2.5.2. One may relax the system class of the minimum phasable system by allowing more capability to the generalized hold; that is, it is supposed that the generalized hold can utilize the state of the zero dynamics $\eta(kT_s)$ by connecting an additional feedback loop. Specifically, the form of the generalized hold that we propose is given by

$$h_g(t - kT_s; \eta(kT_s)) := \begin{cases} h_1(\eta(kT_s)), & kT_s \le t < kT_s + \frac{T_s}{2}, \\ h_2(\eta(kT_s)), & kT_s + \frac{T_s}{2} \le t < kT_s + T_s \end{cases}$$

where $h_1(\eta(kT_s)) + h_2(\eta(kT_s)) = 2$. With this, the zero dynamics under fast sampling (2.5.13) is altered as

$$\begin{aligned} \delta\eta(kT_s) &= C(0, \eta(kT_s)) - \frac{\psi \bar{h}_1(\eta(kT_s))\phi(0, \eta(kT_s))}{4\psi + \frac{\partial \phi}{\partial \xi_1}(0, \eta(kT_s))\psi \bar{h}_1(\eta(kT_s))} \frac{\partial C}{\partial \xi_1}(0, \eta(kT_s)) \\ &=: \gamma(\eta(kT_s)) \end{aligned}$$

where $\bar{h}_i(\eta(kT_s)) = h_i(\eta(kT_s))/T_s, i = 1, 2$. Now, since the gain $\bar{h}_1(\eta(kT_s))$ becomes an arbitrary function of $\eta(kT_s)$, the modified zero dynamics, $\delta\eta(kT_s) = \gamma(\eta(kT_s))$, secures much more flexibility than before, which implies that the class of the minimum phasable system is relaxed. But still, it is not always possible to make the zero dynamics, $\delta\eta(kT_s) = \gamma(\eta(kT_s))$, become stable. When the desired function $\gamma(\eta(kT_s))$ is decided, it should be carefully checked that the hold gain $\bar{h}_1(\eta(kT_s))$ is realizable or not. This can be confirmed by checking the denomina-

tor of the following equation becomes zero or not.

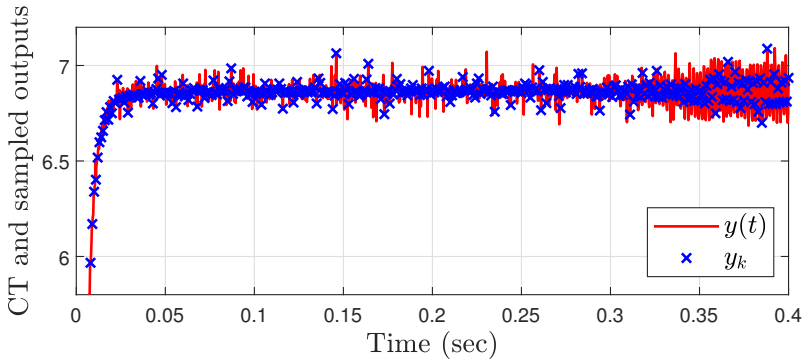
$$\bar{h}_1(\eta(kT_s)) = \frac{4\big(C(0,\eta(kT_s)) - \gamma(\eta(kT_s))\big)}{\phi(0,\eta(kT_s))\frac{\partial C}{\partial \xi_1}(0,\eta(kT_s)) - \frac{\partial \phi}{\partial \xi_1}(0,\eta(kT_s))\big(C(0,\eta(kT_s)) - \gamma(\eta(kT_s))\big)}.$$

Meanwhile, for the case that the given system is not possible to modify the system $\delta\eta(kT_s) = \gamma(\eta(kT_s))$ becomes globally asymptotically stable with any $\bar{h}_1(\eta(kT_s))$, it can be detoured the situation by guaranteeing local stability of the zero dynamics and restrict the region of interest. □

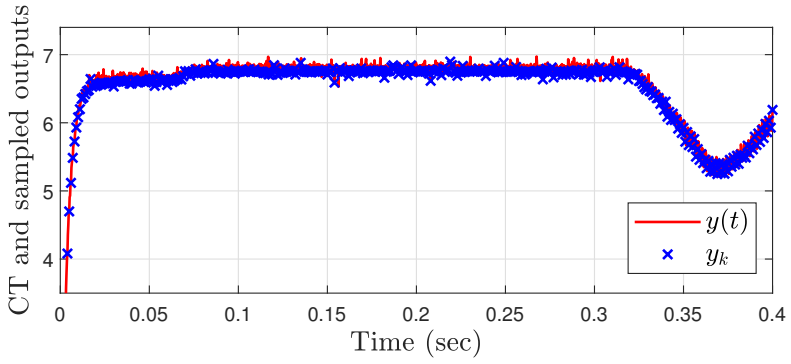
Remark 2.5.3. When the sampled-data system (2.5.11) became minimum phase using GH, one may have wonder regarding to design a controller for the resultant system:

$$\begin{aligned}
\delta\xi_1(kT_s) &= \phi(\xi_1(kT_s),\eta(kT_s)) + \psi u_k + \frac{1}{2^2}\frac{\partial \phi}{\partial \xi_1}\psi\bar{h}_1 u_k, \\
&=: \phi(\xi_1(kT_s),\eta(kT_s)) + \bar{b}(\xi(kT_s),\eta(kT_s)))u_k, \\
\delta\eta(kT_s) &= C(\xi_1(kT_s),\eta(kT_s)) + \frac{1}{2^2}\frac{\partial C}{\partial \xi_1}\psi\bar{h}_1 u_k. \\
&=: C(\xi_1(kT_s),\eta(kT_s)) + \bar{D}(\xi(kT_s),\eta(kT_s))u_k
\end{aligned}$$

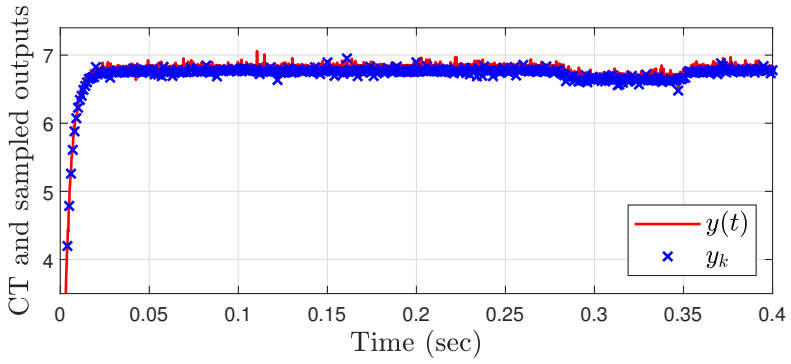
□



(a)



(b)



(c)

Figure 2.16: Experiment results for: (a) Zero dynamics attack $a^{\mathsf{a}}(t)$. (b) Continuous-time output $y(t)$ and sampled output y_k when the zero dynamics attack (2.4.7) is injected into the system. (c) Continuous-time output $y(t)$ and sampled output y_k when the zero dynamics attack constructed with the altered zeros is injected into the system.

Chapter 3

Use of Generalized Hold Feedback in Sampled-data Systems to Counteract Zero-dynamics Sensor Attack

3.1 Undetectable Sensor Attack and its lethality

As we discussed in Section 1.3, a classical method to reveal (or detect) malicious sensor attacks is to equip the control system with the anomaly detector, which assesses whether the attack exists or not by checking a residual signal that is defined as the difference of the measured output and the estimated output. When the size of the residual becomes larger than a predefined threshold (usually, the threshold is set to be a reasonable value so that it does not respond to the sensor noise), it raises the alarm. These anomaly detectors can easily detect simple sensor faults or attacks.

Unfortunately, however, there exists a sensor attack that is undetectable from the anomaly detector (1.3.3). This is the case when the zero dynamics attack is constructed by utilizing the knowledge of the whole system, and it is possible to inject the attack through the sensor network. We call this *zero dynamics sensor attack* as stated before. In what follows, we show how the zero dynamics sensor attack can be constructed in the sampled-data system framework and verify the effect of the attack with an example.

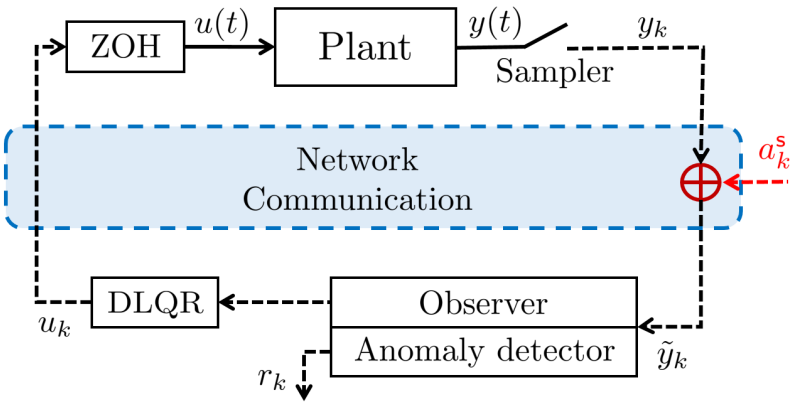


Figure 3.1: Sampled-data control system with sensor attack

3.1.1 Construction of Zero Dynamics Sensor Attack

We start with a controllable and observable continuous-time SISO linear system given by

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t), \end{aligned} \tag{3.1.1}$$

where $x \in \mathbb{R}^n$ is the state vector, $u \in \mathbb{R}$ is the control input, $y \in \mathbb{R}$ is the system output, and A, B, and C are constant matrices with appropriate dimensions.

Similar to the Chapter 2, it is supposed that the system (3.1.1) is controlled by a digital controller through the network communication, and the sampling times of the ZOH and sampler are the same, namely, T_s. In particular, the digital controller is composed of the state estimator, state feedback, and anomaly detector. It is also assumed that the output network is compromised by an attacker who can inject a malicious sensor attack a_k^{s} into the sensor network, which implies that the digital controller receives a contaminated output $\tilde{y}_k := y_k + a_k^{\mathsf{s}}$ rather than y_k.

Now, let us consider the compromised sampled-data system with a digital controller as depicted in Fig. 3.1. With the ZOH, sampler and compromised

sensor network, the sampled-data system of (3.1.1) becomes

$$\begin{aligned} x_{k+1} &= A_{\mathsf{d}} x_k + B_{\mathsf{d}} u_k, \\ \tilde{y}_k &= C_{\mathsf{d}} x_k + a_k^{\mathsf{s}}, \end{aligned} \tag{3.1.2}$$

where $x_k = x(kT_s)$, $k = 0, 1, \ldots,$ is the state vector, u_k and a_k^{s} are the input and the sensor attack, respectively, and $A_{\mathsf{d}} := e^{AT_s}, B_{\mathsf{d}} := \int_0^{T_s} e^{A(T_s-\tau)} B d\tau$, and $C_{\mathsf{d}} := C$. It is assumed that the sampled-data system (3.1.2) is also controllable and observable. Subsequently, the digital controller consists of the Luenberger observer, anomaly detector, and discrete-time linear quadratic regulator (DLQR) [KS72], which are given as follows:

$$\hat{x}_{k+1} = A_{\mathsf{d}} \hat{x}_k + B_{\mathsf{d}} u_k + L(\tilde{y}_k - C_{\mathsf{d}} \hat{x}_k), \tag{3.1.3a}$$

$$r_k = \tilde{y}_k - C_{\mathsf{d}} \hat{x}_k, \tag{3.1.3b}$$

$$u_k = K \hat{x}_k, \tag{3.1.3c}$$

where L and K are selected to guarantee the Schur stability of $A_{\mathsf{d}} - LC_{\mathsf{d}}$ and $A_{\mathsf{d}} + B_{\mathsf{d}} K$, respectively. In particular, the gain K is chosen such that the DLQR $u_k = K\hat{x}_k$ minimizes the performance index

$$J_{\mathsf{d}}(u_k) = \sum_{k=0}^{\infty} (\hat{x}_k^\top Q_{\mathsf{d}} \hat{x}_k + u_k^\top R_{\mathsf{d}} u_k), \tag{3.1.4}$$

where R_{d} is a positive definite matrix and Q_{d} is a non-negative definite matrix such that $(Q_{\mathsf{d}}, A_{\mathsf{d}})$ is detectable. Specifically, the K is computed by

$$K = -(B_{\mathsf{d}}^\top S B_{\mathsf{d}} + R_{\mathsf{d}})^{-1} B_{\mathsf{d}}^\top S A_{\mathsf{d}}$$

with the positive definite matrix S being the solution to the discrete-time Riccati equation

$$A_{\mathsf{d}}^\top S A_{\mathsf{d}} - S - A_{\mathsf{d}}^\top S B_{\mathsf{d}} (B_{\mathsf{d}}^\top S B_{\mathsf{d}} + R_{\mathsf{d}})^{-1} B_{\mathsf{d}}^\top S A_{\mathsf{d}} + Q_{\mathsf{d}} = 0.$$

Now, in order to see the relation between the residual r_k and the sensor

attack a_k^{s} explicitly, we define an error variable $e_k := \hat{x}_k - x_k$, with which the error dynamics can be written as

$$\begin{aligned} e_{k+1} &= (A_{\mathsf{d}} - LC_{\mathsf{d}})e_k + La_k^{\mathsf{s}}, \\ r_k &= -C_{\mathsf{d}}e_k + a_k^{\mathsf{s}} \end{aligned} \tag{3.1.5}$$

(in fact, this is a discrete-time version of the case 1 of Section 1.3). Note that if there is no attack (i.e., $a_k^{\mathsf{s}} \equiv 0$), the residual r_k goes to zero asymptotically (since $A_{\mathsf{d}} - LC_{\mathsf{d}}$ is stable), whereas if there exists an attack, i.e., $a_k^{\mathsf{s}} \neq 0$, the residual r_k directly reacts to the attack so that it can be easily detected.

However, in the attacker's view point, a_k^{s} and r_k are regarded as the input and output of the system (3.1.5), respectively. Hence, finding the zero dynamics of (3.1.5), the attacker can design an undetectable sensor attack (*zero dynamics sensor attack*) a_k^{s} by using the policy given in [TSSJ15]. It is noticed that when the input a_k^{s} of (3.1.5) is set as $a_k^{\mathsf{s}} = C_{\mathsf{d}}e_k$, the output becomes zero (i.e., $r_k = 0$), and in this case, the internal dynamics (zero dynamics) is given by

$$e_{k+1} = A_{\mathsf{d}}e_k, \tag{3.1.6}$$

which is the zero dynamics of (3.1.5) with the output r_k and the input a_k^{s}. With this, the zero dynamics sensor attack can be generated as follows.

Proposition 3.1.1. It is supposed that the sampled-data system equipped with anomaly detector is given by (3.1.1), (3.1.2), and (3.1.3a), whose continuous-time plant (3.1.1) is unstable. Suppose that a sensor attack a_k^{s} is generated by

$$\begin{aligned} x_{k+1}^H &= A_H x_k^H, \\ a_k^{\mathsf{s}} &= C_{\mathsf{d}} x_k^H, \end{aligned} \tag{3.1.7}$$

where A_H is the copy of the zero dynamics of the error dynamics (3.1.5), and the initial condition x_0^H is non-zero (but sufficiently small) and does not belong to the stable eigenspace of A_H. Then, the attack a_k^{s} is stealthy from the anomaly detector, while it makes the states of continuous-time plant (3.1.1) diverge. □

Proof: It is firstly noted that since the zero dynamics of (3.1.5) is (3.1.6),

$A_H = A_{\mathsf{d}}$, and the hacker's state x^H goes to infinity because A_{d} is unstable (this is obtained from the fact that the instability of the matrix A implies that of A_{d} [YG14]), and x_0^H triggers the unstable mode of A_{d}. Now, with a new error variable $\tilde{e}_k := e_k - x_k^H$, the error dynamics with a_k^{s} plugged into (3.1.5) is written by

$$\begin{aligned}\tilde{e}_{k+1} &= A_{\mathsf{d}}\tilde{e}_k + L(-C_{\mathsf{d}}e_k + a_k^{\mathsf{s}}) = (A_{\mathsf{d}} - LC_{\mathsf{d}})\tilde{e}_k,\\ r_k &= -C_{\mathsf{d}}e_k + a_k^{\mathsf{s}} = -C_{\mathsf{d}}\tilde{e}_k.\end{aligned}$$

Then, since $A_{\mathsf{d}} - LC_{\mathsf{d}}$ is stable, one can see that $\|\tilde{e}_k\| \le \alpha z^k \|\tilde{e}_0\|$, where α and z are positive constants with $0 < z < 1$. This means that $\tilde{e}_k$ converges to zero, and so does the residual r_k. Hence, if the initial condition x_0^H is small enough, the residual r_k is kept lower than the threshold for all k, which shows that the attack a_k^{s} is undetectable.

Meanwhile, when the attack a_k^{s} is plugged into (3.1.3a), its closed loop system becomes

$$\hat{x}_{k+1} = (A_{\mathsf{d}} + B_{\mathsf{d}}K)\hat{x}_k - LC_{\mathsf{d}}(\hat{x}_k - x_k - x_k^H).$$

From the fact that $\hat{x}_k - x_k - x_k^H (= \tilde{e}_k)$ converges to zero and $A_{\mathsf{d}} + B_{\mathsf{d}}K$ is stable, we know that the state of the observer $\hat{x}_k$ goes to zero asymptotically. Consequently, since $\tilde{e}_k \to 0$ as $k \to \infty$ implies that the error variable e $(= \hat{x} - x)$ follows the attacker's state trajectory x^H, we know that the state $-x_k$ diverges to follow the attacker's state x_k^H, which completes the proof. ■

3.1.2 Simulation Results: Magnetic Levitation of a Steel Ball

Consider a magnetic levitation system [YT01] shown in Fig. 3.2, which works as follows. When the current flows through the coil, the electromagnetic force is generated, and it makes the steel ball levitate. Meanwhile, the laser sensor measures the position of the steel ball, and then it is sent to the computer through the A/D converter. Then, the computer provides a discrete control signal, by which the current of the coil is controlled to levitate the steel ball at a fixed position (i.e., equilibrium point).

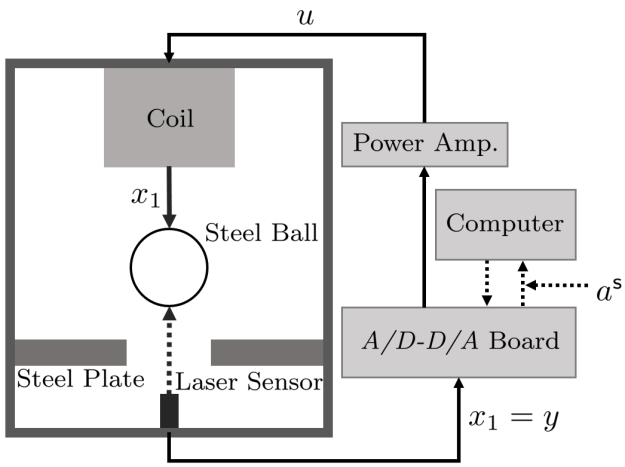


Figure 3.2: Magnetic levitation system under sensor attack

A linearized magnetic levitation system (at the equilibrium) is given by

$$\begin{aligned} \dot{x}(t) &= \begin{bmatrix} 0 & 1 \\ 3270 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ -26.67 \end{bmatrix} u(t), \\ y(t) &= \begin{bmatrix} 1 & 0 \end{bmatrix} x(t), \end{aligned} \tag{3.1.8}$$

where $x := [x_1 \ x_2]^\top$; x_1 and x_2 are the position and velocity of the steel ball, respectively, and u is the current. The corresponding sampled-data system with the sampling period $T_s = 0.01$ sec becomes

$$\begin{aligned} x_{k+1} &= \begin{bmatrix} 1.17 & 0.01 \\ 34.51 & 1.17 \end{bmatrix} x_k + \begin{bmatrix} -0.001 \\ -0.28 \end{bmatrix} u_k, \\ y_k &= \begin{bmatrix} 1 & 0 \end{bmatrix} x_k. \end{aligned} \tag{3.1.9}$$

In the digital controller (3.1.3a), the gains of the observer and DLQR are chosen as $L = [2.04 \ \ 132.47]^\top$ and $K = [126.14 \ \ 4.17]$ with

$$Q_{\mathsf{d}} = \begin{bmatrix} 1000 & 0 \\ 0 & 1005 \end{bmatrix} \quad \text{and} \quad R_{\mathsf{d}} = 0.0001.$$

Subsequently, it is assumed that the discretized sensor measurement y_k is

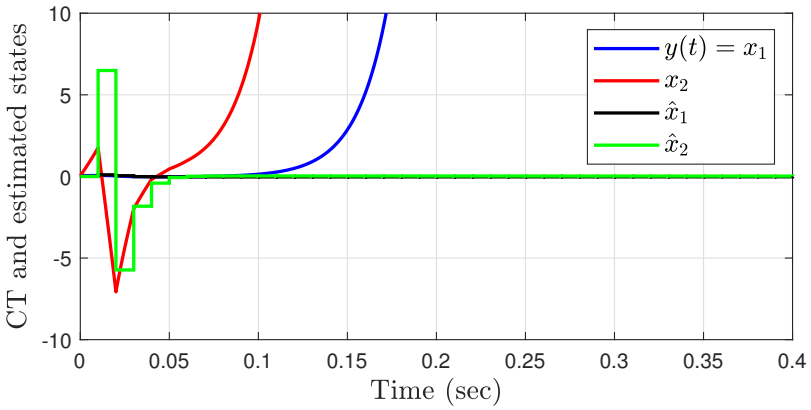


Figure 3.3: Continuous-time state trajectories of the plant (blue and red lines) and its discretized estimations (black and green lines).

contaminated by a malicious attacker while it is delivered to the estimator, which is generated by Proposition 3.1.1. That is,

$$x_{k+1}^H = \begin{bmatrix} 1.17 & 0.01 \\ 34.51 & 1.17 \end{bmatrix} x_k^H, \quad a_k^{\mathsf{s}} = \begin{bmatrix} 1 & 0 \end{bmatrix} x_k^H,$$

where $x_0^H = 10^{-2} \times [-0.1 \ \ -0.5]^\top$, and a_k^{s} is injected into the system at 0 sec. Then, one can see in Fig. 3.3 that the continuous-time states of the physical plant are spoiled because of the effect of the attack, whereas, as discussed in the proof of Proposition 3.1.1, the states of the observer converge to zero. Meanwhile, as seen in Fig. 3.4, the residual r_k is maintained at almost zero even though the size of the attack is increased, which means the injected sensor attack a_k^{s} is hardly detectable.

3.2 Strategy to Neutralize Zero Dynamics Sensor Attack and Relieve Performance Degradation

As seen in the previous section, the zero dynamics sensor attack is powerful, because it is hard to detect, and it makes the system states go to the unsafe region while the physical plant is unstable. Noting the fact that whether the attack is

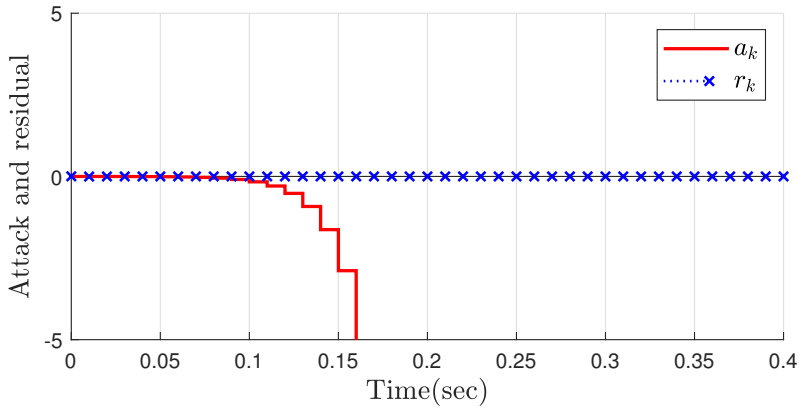


Figure 3.4: Injected sensor attack (red line) and the residual (blue cross) of the anomaly detector.

effective or not depends on the stability of the hacker's dynamics (3.1.7), which is a copy of the system matrix of the sampled-data system (3.1.2), we focus on to incapacitate the attack itself by stabilizing the sampled-data system rather than detecting the attack directly.

To realize this idea, we employ the generalized hold again. But this time, we use a generalized hold with an output feedback loop as seen in Fig. 3.5, and by doing so, the poles of the sampled-data system (3.1.2) can be assigned arbitrary locations. In what follows, we first show that the zero dynamics sensor attack (3.1.1) can be neutralized by shifting the poles of the sampled-data system (3.1.2) into the inside of the unit circle. Then, in order to relieve performance degradation of the use of generalized hold, we propose to use DLQR that minimizes performance index composed of continuous-time rather than the discrete-time one (3.1.4). Finally, we close this chapter with a demonstration of the proposed strategy using the magnetic levitation system.

3.2.1 Employing the generalized hold feedback to neutralize zero dynamics sensor attack

We start with recalling the "generalized hold", which is a signal holding device that converts a discrete-time signal v_k into a continuous one $v(t)$ using a predefined

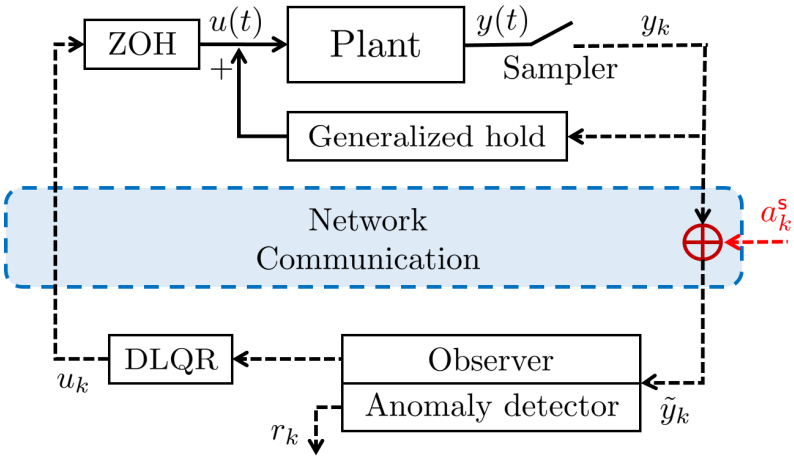


Figure 3.5: Sampled-data control system with generalized hold feedback

function $f_g(t)$ defined on $[0, T_s)$; that is,

$$v(t) = f_g(t - kT_s)v_k, \quad kT_s \le t < (k+1)T_s.$$

(If $f_g(t) = 1$ for $0 \le t < T_s$, this is noting but the ZOH.) When the generalized hold is utilized for a feedback as seen in Fig 3.5, the input of the plant becomes

$$u(kT_s + t) = u_k + f_g(t)y_k, \quad 0 \le t < T_s,$$

with which the sampled-data system from the input u_k to the output $\tilde{y}_k$ is obtained as follows:

$$\begin{aligned} x_{k+1} &= e^{AT_s}x_k + \int_0^{T_s} e^{A(T_s-\tau)}Bd\tau u_k + \int_0^{T_s} e^{A(T_s-\tau)}Bf_g(\tau)d\tau y_k, \\ \tilde{y}_k &= C_{\mathsf{d}}x_k + a_k^{\mathsf{s}}, \end{aligned}$$

and it can be rewritten as

$$\begin{aligned} x_{k+1} &= (A_{\mathsf{d}} + F_gC_{\mathsf{d}})x_k + B_{\mathsf{d}}u_k, \\ \tilde{y}_k &= C_{\mathsf{d}}x_k + a_k^{\mathsf{s}}, \end{aligned} \tag{3.2.1}$$

where

$$F_g := \int_0^{T_s} e^{A(T_s-\tau)} B f_g(\tau) d\tau. \tag{3.2.2}$$

It is well known that, due to the controllability of (A, B), the vector $F_g \in \mathbb{R}^n$ can be assigned arbitrarily by finding an appropriate impulse function $f_g(t)$. Here, we postpone how to find the impulse function $f_g(t)$ for a while. Instead, let us keep in mind that F_g can be an arbitrary vector.

It is noted that since the sampled-data system is altered from (3.1.2) to (3.2.1), the corresponding digital controller is needed to be redesigned as the followings:

$$\begin{aligned} \hat{x}_{k+1} &= (A_{\mathsf{d}} + F_g C_{\mathsf{d}})\hat{x}_k + B_{\mathsf{d}} u_k + \bar{L}(\tilde{y}_k - C_{\mathsf{d}}\hat{x}_k), \\ r_k &= \tilde{y}_k - C_{\mathsf{d}}\hat{x}_k, \\ u_k &= \bar{K}\hat{x}_k, \end{aligned} \tag{3.2.3}$$

where $\bar{L}$ and $\bar{K}$ are new gains of the observer and DLQR, respectively. Again, in order to investigate the effect of the zero dynamics sensor attack on the altered system (3.2.1), we find a zero dynamics of the error dynamics (with an error variable $e_k = \hat{x}_k - x_k$). This can be obtained as follows:

$$\begin{aligned} e_{k+1} &= (A_{\mathsf{d}} + F_g C_{\mathsf{d}} - \bar{L}C_{\mathsf{d}})e_k + \bar{L}a_k^{\mathsf{s}}, \\ r_k &= -C_{\mathsf{d}} e_k + a_k^{\mathsf{s}}, \end{aligned}$$

and its zero dynamics is given by

$$e_{k+1} = (A_{\mathsf{d}} + F_g C_{\mathsf{d}})e_k. \tag{3.2.4}$$

We already knew from Proposition 3.1.1 that the effectiveness of the zero dynamics sensor attack depends heavily on the stability of the zero dynamics of the error dynamics (see Eq. (3.1.7) and (3.2.4)). Thus, if we choose F_g to make the matrix $A_{\mathsf{d}} + F_g C_{\mathsf{d}}$ be Schur stable, then the zero dynamics sensor attack becomes ineffective even if it is still undetectable.

On the other hand, the attacker who does not recognize the existence of

the generalized hold feedback may construct the attack signal using the old zero dynamics (3.1.6), not the new one (3.2.4). In this case, of course, the attack is effective, but it is not stealthy anymore.

Remark 3.2.1. In this section, we suggest employing the generalized hold feedback to make the discrete-time system become stable to neutralize the zero dynamics sensor attack. However, this is useful only when the system is not output feedback stabilizable. If it is, the generalized hold $f_g(t)$ is not needed since with static high gain output feedback, the system can be stabilized. Suppose there is a static output feedback from y_k to $u(t)$ with gain g_{s}. Then, the system (3.2.1) becomes

$$\begin{aligned} x_{k+1} &= (A_{\mathsf{d}} + g_{\mathrm{s}} B_{\mathsf{d}} C_{\mathsf{d}}) x_k + B_{\mathsf{d}} u_k \\ \tilde{y}_k &= C_{\mathsf{d}} x_k + a_k^{\mathsf{s}}, \end{aligned}$$

and its transfer function is given by

$$\frac{N(z)}{D(z) - g_{\mathrm{s}} N(z)}$$

where $N(z)/D(z) := C_{\mathsf{d}}(zI_n - A_{\mathsf{d}})^{-1} B_{\mathsf{d}}$. One sufficient condition for output stabilizable is that the system $(A_{\mathsf{d}}, B_{\mathsf{d}}, C_{\mathsf{d}})$ is minimum phase because as g_{s} becomes larger, poles of the system close to roots of $N(z)$, which implies the minimum phase system can be stabilized by high gain output feedback. □

Remark 3.2.2. Once the F_g is selected to stabilize $A_{\mathsf{d}} + F_g C_{\mathsf{d}}$, the gains $\bar{L}$ and $\bar{K}$ may be redesigned. First, to properly estimate the state of the altered system (3.2.1), $\bar{L}$ should be re-chosen to stabilize the matrix $A_{\mathsf{d}} + (F_g - \bar{L}) C_{\mathsf{d}}$. Meanwhile, since the system (3.2.1) is already stable now, $\bar{K}$ may be regarded as an outer loop control gain, so it is focused on to increase other control performances (e.g., optimal control) as long as $A_{\mathsf{d}} + F_g C_{\mathsf{d}} + B_{\mathsf{d}} \bar{K}$ is stable. However, in some cases, it is possible to use the original values; i.e., $\bar{K} = K$ or $\bar{L} = L$. □

Now, we back to the problem of finding $f_g(t)$ when an appropriate F_g is determined (to protect the system from the zero dynamics sensor attack). As a matter of fact, the procedure of finding $f_g(t)$ is the same as Lemma (2.2.2), but for

the completion of the dissertation, we introduce this once again in the following. When the desired F_g is denoted by F_g^*, one can readily find $f_g(t)$ that satisfies (3.2.2) by the following function:

$$f_g(t) = B^\top e^{A^\top (T_s - t)} W^{-1}(0, T_s) F_g^*, \tag{3.2.5}$$

where $W(0, T_s)$ is the controllability Gramian.

On the other hand, in some cases, the continuous function (3.2.5) may not be suitable for someone who wants to implement the generalized hold $f_g(t)$ in a digital device. In this case, one can use a piecewise constant function $f_g(t)$ instead of (3.2.5). Suppose $f_g(t)$ is a piecewise constant with N subintervals, i.e.,

$$f_g(t) = f_i, \ \frac{(i-1)T_s}{N} \le t < \frac{iT_s}{N}, \ i = 1, \dots, N. \tag{3.2.6}$$

By substituting (3.2.6) into (3.2.2), one has

$$F_g = \sum_{l=1}^{N} f_l \int_{\frac{(l-1)T_s}{N}}^{\frac{lT_s}{N}} e^{A(T_s - \tau)} B d\tau,$$

which can be rewritten as

$$F_g = \begin{bmatrix} A_{\mathsf{d},N}^{N-1} B_{\mathsf{d},N} & \cdots & A_{\mathsf{d},N} B_{\mathsf{d},N} & B_{\mathsf{d},N} \end{bmatrix} f = \mathcal{C}_{\mathsf{d},N} f,$$

where $f := [f_1 \ \cdots \ f_N]^\top$, $A_{\mathsf{d},N} = e^{A\frac{T_s}{N}}$, $B_{\mathsf{d},N} = \int_0^{T_s/N} e^{A(\frac{T_s}{N} - \tau)} B d\tau$, and $\mathcal{C}_{\mathsf{d},N} \in \mathbb{R}^{n \times N}$. Then, under the assumption that $(A_{\mathsf{d},N}, B_{\mathsf{d},N})$ is controllable, the gain f of the piecewise constant function $f_g(t)$ with $N \ge n$ can be readily obtained by

$$f = \mathcal{C}_{\mathsf{d},N}^\dagger F_g^*. \tag{3.2.7}$$

Hence, as aforementioned, one can assign the eigenvalues of $A_\mathsf{d} + F_g C_\mathsf{d}$ at desired locations by appropriate F_g obtained from (3.2.2) with (3.2.5) or (3.2.6).

3.2.2 Simulation Results: Effectiveness of the Generalized Hold

Let us reconsider the magnetic levitation system equipped with the generalized hold feedback. The continuous-time plant under consideration is the same as that of Section 3.1.2 (Eq. (3.1.8)); that is,

$$x_{k+1} = \begin{bmatrix} 1.17 & 0.01 \\ 34.51 & 1.17 \end{bmatrix} x_k + \begin{bmatrix} -0.001 \\ -0.28 \end{bmatrix} u_k,$$
$$y_k = \begin{bmatrix} 1 & 0 \end{bmatrix} x_k.$$

Firstly, we select F_g (3.2.2) by $F_g = [-0.45 \;\; -39.03]^\top$, which shifts the eigenvalues of $A_\mathsf{d} + F_g C_\mathsf{d}$ to the stable region; 0.9 and 0.99. Then, the obtained F_g is realized by using the piecewise constant function $f_g(t)$ of (3.2.6) and (3.2.7) with $N = 2$ subintervals. By doing so, the generalized hold gain is obtained by $f = [520.01 \;\; -274.1]^\top$ (Of course, one can use the controllability Gramian, (3.2.5) to achieve the impulse function $f_g(t)$). As we mentioned in Remark 3.2.2, after we design the generalized hold feedback, it should be checked whether the old gain L that used in Section 3.1.2 is still valid or not, because the system matrix of the sampled-data system is changed from A_d to $A_f = A_\mathsf{d} + F_g C_\mathsf{d}$ (i.e., $A_\mathsf{d} + (F_g - L)C_\mathsf{d}$ is stable or not). In this case, $A_\mathsf{d} + (F_g - L)C_\mathsf{d}$ is unstable with the predefined ones, so we redesign the gain of the observer by $\bar{L} = [0.99 \;\; 44.09]^\top$. Meanwhile, we redesign the control gain (DLQR) $\bar{K}$ with the altered dynamics (3.2.1) and performance index (3.1.4). The obtained control gain is denoted $\bar{K}_\mathsf{d}$, which is given by

$$\bar{K}_\mathsf{d} = [-16.01 \;\; 4.15] \tag{3.2.8}$$

with

$$Q_\mathsf{d} = \begin{bmatrix} 1000 & 0 \\ 0 & 1005 \end{bmatrix} \quad \text{and} \quad R_\mathsf{d} = 0.0001.$$

Now, in order to investigate the security enhancement of the proposed protection strategy, we consider two different attacking scenarios. The first one is that the attacker, who does not notice the generalized hold feedback is supplemented,

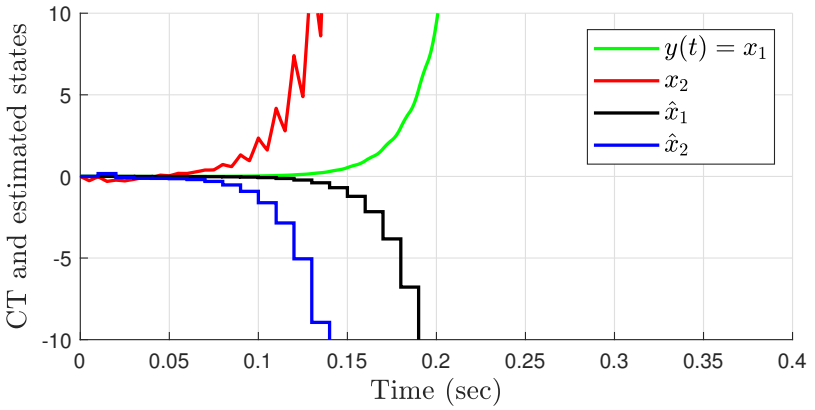


Figure 3.6: Continuous-time state trajectories of the plant (green and red lines) and its discretized estimations (black and blue lines) with the generalized hold feedback and the attack in Section 3.1.2

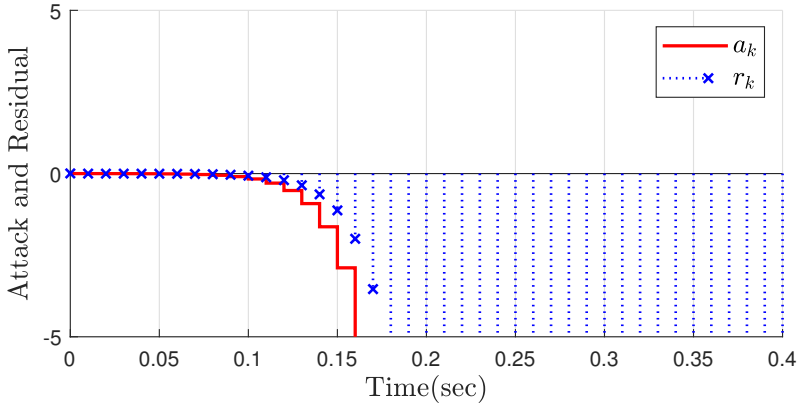


Figure 3.7: Injected sensor attack without consideration of the generalized hold feedback (red line) and the residual of the anomaly detector (blue cross).

injects the same attack that of Section 3.1.2. In this case, the attack is still dangerous, as in Section 3.1.2. However, the zero dynamics of the error dynamics is not the same anymore (the zero dynamics is altered from (3.1.6) to (3.2.4)) so that the injected attack is not stealthy anymore and its harm can be captured from the anomaly detector as one can see in Fig. 3.7.

As the second scenario, let us suppose that a clever attacker knows not only the fact that the system is equipped with the generalized hold feedback but also

the exact hold gain $f_g(t)$. Thus, the attacker reconstructs the attack (3.1.7) with the new zero dynamics (3.2.4), and they injects the attack into the system at 0 sec. The injected sensor attack and the residual of the anomaly detector are illustrated in Fig. 3.8. Although the injected attack is undetectable again, this is nothing but a vanishing perturbation since $A_{\mathsf{d}} + F_g C_{\mathsf{d}}$ is stable and the initial condition of (3.1.7) is sufficiently small. Thus, it is easily expected that the system states hardly affected by this attack so that they converge to the origin. This can be seen in Fig. 3.9.

3.2.3 DLQR under Consideration of Inter-sample Behavior

As we discussed in Section 2.2.2, because of its nature, the generalized hold is bound to cause fluctuations in the inter-sample behavior of the physical plant even if it is used as the output feedback form. Therefore, in this section, we would like to construct a digital control law that alleviates undesirable fluctuations in the inter-sample behavior. To this end, we consider a discrete-time linear quadratic regulator problem, in which the discrete control u_k minimizes a performance index consisting of the continuous-time rather than the sampled one (see. (3.1.4)). As a matter of fact, this concept has already been introduced in [FG96], but we suggest its use when the generalized hold feedback loop exists as in Fig. 3.5.

Let us suppose that the continuous-time plant (3.1.1) is equipped with the generalized hold feedback designed by the procedure described in Section 3.2.1. Then, the resulting continuous-time input $u(t)$ is written by

$$u(t) = u_k + f_g(t - kT_s)y_k, \quad kT_s \le t < kT_s + T_s, \tag{3.2.9}$$

and then the continuous-time plant (3.1.1) becomes

$$\dot{x}(t) = Ax(t) + Bf_g(t - kT_s)Cx(kT_s) + Bu_k, \quad kT_s \le t < kT_s + T_s. \tag{3.2.10}$$

Under this setting, we would like to find a discrete-time control law u_k that minimizes the following continuous-time performance index:

$$J_{\mathsf{c}}(u) = \int_0^\infty \big(x(t)^\top Q_{\mathsf{c}} x(t) + u(t)^\top R_{\mathsf{c}} u(t)\big)dt. \tag{3.2.11}$$

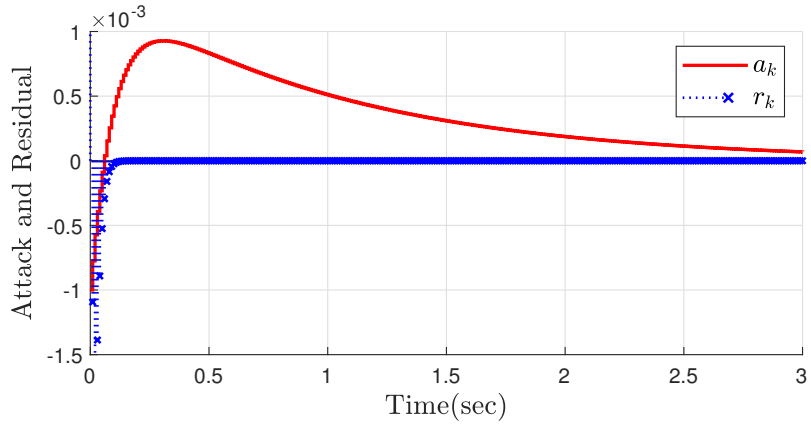


Figure 3.8: Injected sensor attack under consideration of the generalized hold feedback (red line) and the residual of the anomaly detector (blue cross).

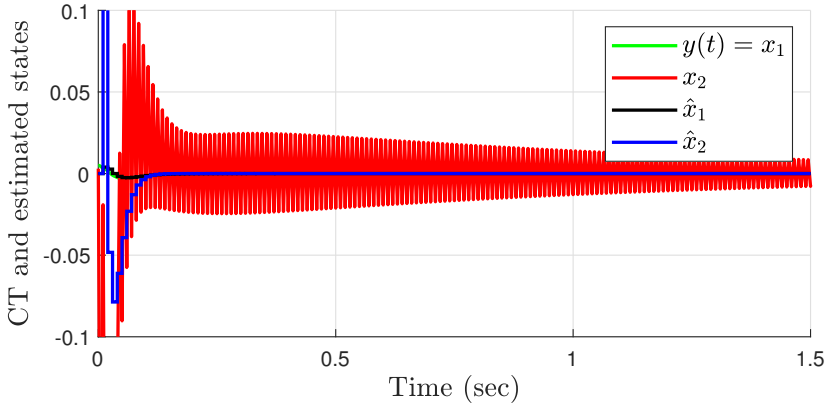


Figure 3.9: Continuous-time state trajectories of the plant (green and red lines) and discrete-time state estimations with the generalized hold feedback (black and blue lines).

It is important to note that the continuous-time performance index J_{c} evaluates the cost including the inter-sample response of the physical plant whereas the discrete-time one, J_{d} (3.1.4), cannot cover the inter-sample behavior of the plant (it focuses only the sampled response). Hence, compared to use the DLQR with the discrete-time performance index J_{d} (3.1.4), it is expected that DLQR with the continuous-time performance index can attenuate the fluctuation of the inter-sample behavior caused by generalized hold feedback.

We now convert the continuous-time cost function (3.2.11) into the discrete-time one to get conventional DLQR form. With the dynamics with generalized hold feedback (3.2.10), for $kT_s \le t < kT_s + T_s$,

$$\begin{aligned} x(t) &= e^{A(t-kT_s)}x(kT_s) \\ &\quad + \int_{kT_s}^{t} e^{A(t-\tau)}Bf_g(\tau - kT_s)Cd\tau x(kT_s) + \int_{kT_s}^{t} e^{A(t-\tau)}Bd\tau u(kT_s), \\ u(t) &= u(kT_s), \end{aligned}$$

and by defining a time-varying matrix, the continuous-time state $x(t)$ and input $u(t)$ can be rewritten by the discrete-time state x_k and input u_k as follows:

$$\begin{aligned} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} &= \begin{bmatrix} e^{A(t-kT_s)} + \int_{kT_s}^{t} e^{A(t-\tau)}Bf_g(\tau - kT_s)Cd\tau & \int_{kT_s}^{t} e^{A(t-\tau)}Bd\tau \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x(kT_s) \\ u(kT_s) \end{bmatrix} \\ &=: \begin{bmatrix} \alpha(t) & \beta(t) \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix}, \end{aligned} \tag{3.2.12}$$

where

$$\alpha(t) := e^{A(t-kT_s)} + \int_{kT_s}^{t} e^{A(t-\tau)}Bf_g(\tau - kT_s)Cd\tau \quad \text{and} \quad \beta(t) := \int_{kT_s}^{t} e^{A(t-\tau)}Bd\tau.$$

Lemma 3.2.1. The continuous-time performance index $J_{\mathsf{c}}(u)$ (3.2.11) can be rewritten as a discrete-time performance index as follows:

$$J_{\mathsf{c}}(u) := \sum_{k=0}^{\infty} [x_k^\top \; u_k^\top] \begin{bmatrix} \bar{Q}_{\mathsf{d}} & \bar{V}_{\mathsf{d}} \\ \bar{V}_{\mathsf{d}}^\top & \bar{R}_{\mathsf{d}} \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix}, \tag{3.2.13}$$

where

$$\begin{aligned}\bar{Q}_{\mathsf{d}} &:= \int_0^{T_s} \bar{\alpha}(t)^\top Q_{\mathsf{c}} \bar{\alpha}(t) dt,\\ \bar{R}_{\mathsf{d}} &:= \int_0^{T_s} (\bar{\beta}(t)^\top Q_{\mathsf{c}} \bar{\beta}(t) + R_{\mathsf{c}}) dt,\\ \bar{V}_{\mathsf{d}} &:= \int_0^{T_s} \bar{\alpha}(t)^\top Q_{\mathsf{c}} \bar{\beta}(t) dt,\end{aligned} \tag{3.2.14}$$

with

$$\bar{\alpha}(t) := e^{At} + \int_0^t e^{A(t-\tau)} B f_g(\tau) C d\tau \quad \text{and} \quad \bar{\beta}(t) := \int_0^t e^{A(t-\tau)} B d\tau. \tag{3.2.15}$$

□

Proof: Substituting (3.2.12) into (3.2.11), one has

$$\begin{aligned}J_{\mathsf{c}}(u) &= \int_0^\infty \big(x(t)^\top Q_{\mathsf{c}} x(t) + u(t)^\top R_{\mathsf{c}} u(t)\big) dt\\ &= \sum_{k=0}^\infty \int_{kT_s}^{kT_s+T_s} \begin{bmatrix} x(t)^\top & u(t)^\top \end{bmatrix} \begin{bmatrix} Q_{\mathsf{c}} & 0 \\ 0 & R_{\mathsf{c}} \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} dt\\ &= \sum_{k=0}^\infty \Bigg\{ \begin{bmatrix} x_k^\top & u_k^\top \end{bmatrix} \int_{kT_s}^{kT_s+T_s} \begin{bmatrix} \alpha(t) & \beta(t) \\ 0 & 1 \end{bmatrix}^\top \begin{bmatrix} Q_{\mathsf{c}} & 0 \\ 0 & R_{\mathsf{c}} \end{bmatrix} \begin{bmatrix} \alpha(t) & \beta(t) \\ 0 & 1 \end{bmatrix} dt \begin{bmatrix} x_k \\ u_k \end{bmatrix} \Bigg\}.\end{aligned} \tag{3.2.16}$$

In what follows, with the change of variable $\bar{t} := t - kT_s$, we make the integral term of (3.2.16) not depend on kT_s; that is,

$$\begin{aligned}&\int_{kT_s}^{kT_s+T_s} \begin{bmatrix} \alpha(t) & \beta(t) \\ 0 & 1 \end{bmatrix}^\top \begin{bmatrix} Q_{\mathsf{c}} & 0 \\ 0 & R_{\mathsf{c}} \end{bmatrix} \begin{bmatrix} \alpha(t) & \beta(t) \\ 0 & 1 \end{bmatrix} dt\\ &= \int_0^{T_s} \begin{bmatrix} \alpha(\bar{t}+kT_s) & \beta(\bar{t}+kT_s) \\ 0 & 1 \end{bmatrix}^\top \begin{bmatrix} Q_{\mathsf{c}} & 0 \\ 0 & R_{\mathsf{c}} \end{bmatrix} \begin{bmatrix} \alpha(\bar{t}+kT_s) & \beta(\bar{t}+kT_s) \\ 0 & 1 \end{bmatrix} d\bar{t}\end{aligned}$$

Subsequently, again with the change of variable $\bar{\tau} := \tau - kT_s$, α and β can be

independent to T_s; that is,

$$\begin{aligned}\alpha(\bar{t}+kT_s) &= e^{A\bar{t}} + \int_{kT_s}^{\bar{t}+kT_s} e^{A(\bar{t}+kT_s-\tau)} B f_g(\tau - kT_s) C d\tau \\ &= e^{A\bar{t}} + \int_0^{\bar{t}} e^{A(\bar{t}-\bar{\tau})} B f_g(\bar{\tau}) C d\bar{\tau} =: \bar{\alpha}(\bar{t}), \qquad (3.2.17)\\ \beta(\bar{t}+kT_s) &= \int_{kT_s}^{\bar{t}+kT_s} e^{A(\bar{t}+kT_s-\tau)} B d\tau = \int_0^{\bar{t}} e^{A(\bar{t}-\bar{\tau})} B d\bar{\tau} =: \bar{\beta}(\bar{t}),\end{aligned}$$

which implies

$$\begin{aligned}J_{\mathsf{c}}(u) &= \sum_{k=0}^{\infty}\left\{ [x_k^\top \ u_k^\top] \int_0^{T_s} \begin{bmatrix} \bar{\alpha}(\bar{t}) & \bar{\beta}(\bar{t}) \\ 0 & 1 \end{bmatrix}^\top \begin{bmatrix} Q_{\mathsf{c}} & 0 \\ 0 & R_{\mathsf{c}} \end{bmatrix} \begin{bmatrix} \bar{\alpha}(\bar{t}) & \bar{\beta}(\bar{t}) \\ 0 & 1 \end{bmatrix} d\bar{t} \begin{bmatrix} x_k \\ u_k \end{bmatrix} \right\}, \\ &= \sum_{k=0}^{\infty}\left\{ [x_k^\top \ u_k^\top] \int_0^{T_s} \begin{bmatrix} \bar{\alpha}(\bar{t})^\top Q_{\mathsf{c}} \bar{\alpha}(\bar{t}) & \bar{\alpha}(\bar{t})^\top Q_{\mathsf{c}} \bar{\beta}(\bar{t}) \\ \bar{\beta}(\bar{t})^\top Q_{\mathsf{c}} \bar{\alpha}(\bar{t}) & \bar{\beta}(\bar{t})^\top Q_{\mathsf{c}} \bar{\beta}(\bar{t}) + R_{\mathsf{c}} \end{bmatrix} d\bar{t} \begin{bmatrix} x_k \\ u_k \end{bmatrix} \right\}, \\ &=: \sum_{k=0}^{\infty} [x_k^\top \ u_k^\top] \begin{bmatrix} \bar{Q}_{\mathsf{d}} & \bar{V}_{\mathsf{d}} \\ \bar{V}_{\mathsf{d}}^\top & \bar{R}_{\mathsf{d}} \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix}. \qquad (3.2.18)\end{aligned}$$

(the dummy variable $\bar{t}$ can be replaced by t). Then, (3.2.18) and (3.2.17) correspond to (3.2.14) and (3.2.15), respectively, which completes the proof. ■

Corollary 3.2.2. If $f_g(t)$ is a piecewise constant function with N subintervals, the continuous-time performance index (3.2.11) is represented by

$$\begin{aligned}J_{\mathsf{c}}(u) &= \int_0^{\infty} \big(x(t)^\top Q_{\mathsf{c}} x(t) + u(t)^\top R_{\mathsf{c}} u(t) \big) dt \\ &= \sum_{k=0}^{\infty} \int_{kT_s}^{kT_s+T_s} \begin{bmatrix} x(t)^\top & u(t)^\top \end{bmatrix} \begin{bmatrix} Q_{\mathsf{c}} & 0 \\ 0 & R_{\mathsf{c}} \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} dt \\ &= \sum_{k=0}^{\infty} \left\{ [x_k^\top \ u_k^\top] \sum_{i=1}^{N} \int_{kT_s+\frac{(i-1)T_s}{N}}^{kT_s+\frac{iT_s}{N}} \begin{bmatrix} \alpha_i(t) & \beta(t) \\ 0 & 1 \end{bmatrix}^\top \begin{bmatrix} Q_{\mathsf{c}} & 0 \\ 0 & R_{\mathsf{c}} \end{bmatrix} \begin{bmatrix} \alpha_i(t) & \beta(t) \\ 0 & 1 \end{bmatrix} dt \begin{bmatrix} x_k \\ u_k \end{bmatrix} \right\}\end{aligned}$$

where

$$\alpha_i(t) = e^{A(t-kT_s)} + \int_{kT_s}^{t} e^{A(t-\tau)} B f_i C d\tau.$$

Then, by similar procedure of the proof of Lemma 3.2.1, one has

$$
\begin{aligned}
&J_{\mathsf{c}}(u)\\
&= \sum_{k=0}^{\infty}\left\{ [x_k^\top \ u_k^\top] \sum_{i=1}^{N} \int_{\frac{(i-1)T_s}{N}}^{\frac{iT_s}{N}} \begin{bmatrix} \hat{\alpha}_i(t) & \beta(t) \\ 0 & 1 \end{bmatrix}^\top \begin{bmatrix} Q_{\mathsf{c}} & 0 \\ 0 & R_{\mathsf{c}} \end{bmatrix} \begin{bmatrix} \hat{\alpha}_i(t) & \beta(t) \\ 0 & 1 \end{bmatrix} dt \begin{bmatrix} x_k \\ u_k \end{bmatrix} \right\}\\
&=: \sum_{k=0}^{\infty}\left\{ [x_k^\top \ u_k^\top] \sum_{i=1}^{N} \int_{\frac{(i-1)T_s}{N}}^{\frac{iT_s}{N}} \begin{bmatrix} \hat{Q}_{\mathsf{c},i} & \hat{V}_{\mathsf{c},i} \\ \hat{V}_{\mathsf{c},i}^\top & \hat{R}_{\mathsf{c},i} \end{bmatrix} dt \begin{bmatrix} x_k \\ u_k \end{bmatrix} \right\}
\end{aligned}
$$

where

$$\hat{\alpha}_i(t) = e^{At} + \int_0^t e^{A(t-\tau)} B f_i C d\tau.$$

□

Next, with defining $A_f := A_{\mathsf{d}} + F_g C_{\mathsf{d}}$, we find DLQR $u_k = \bar{K}\hat{x}_k$ minimizing the performance index $J_{\mathsf{c}}(u_k)$ (3.2.18), whose gain $\bar{K}$ is computed by

$$\bar{K} = -(B_{\mathsf{d}}^\top S B_{\mathsf{d}} + \bar{R}_{\mathsf{d}})^{-1}(B_{\mathsf{d}}^\top S A_f + \bar{V}_{\mathsf{d}}^\top) \tag{3.2.19}$$

with the positive definite matrix S being the solution to the discrete-time Riccati equation

$$A_f^\top S A_f - S - (A_f^\top S B_{\mathsf{d}} + \bar{V}_{\mathsf{d}})(B_{\mathsf{d}}^\top S B_{\mathsf{d}} + \bar{R}_{\mathsf{d}})^{-1}(B_{\mathsf{d}}^\top S A_f + \bar{V}_{\mathsf{d}}^\top) + \bar{Q}_{\mathsf{d}} = 0.$$

In particular, the following conditions should be satisfied to solve the Riccati equation:

- (A_f, B_{d}) is stabilizable,
- $(\bar{Q}_{\mathsf{d}} - \bar{V}_{\mathsf{d}}\bar{R}_{\mathsf{d}}^{-1}\bar{V}^\top, A_f - B_{\mathsf{d}}\bar{R}_{\mathsf{d}}^{-1}\bar{V}_{\mathsf{d}}^\top)$ is detectable,
- $\bar{R}_{\mathsf{d}} > 0$ and $\bar{Q}_{\mathsf{d}} - \bar{V}_{\mathsf{d}}\bar{R}_{\mathsf{d}}^{-1}\bar{V}_{\mathsf{d}} \geq 0$.

In contrast with J_{d}, in the continuous-time performance index J_{c}, the state and input cross-term $\bar{V}_{\mathsf{d}}$ exists. By minimizing this input and state correlation (the inter-sample behavior), DLQR, $u_k = \bar{K}\hat{x}_k$, attenuates the fluctuations in the inter-sample behavior.

3.2.4 Simulation Results: Effectiveness of DLQR with Continuous-time Performance Index

As we discussed in the previous subsection, when we employ the generalized hold feedback, due to its nature, the inter-sample behavior has no choice but to fluctuate as seen in Fig. 3.9. In order to attenuate this phenomenon, we use DLQR with the continuous-time performance index (3.2.11) for the magnetic levitation system equipped with the generalized hold feedback. With the security enhanced dynamics (3.2.1) and continuous-time performance index (3.2.11), the control gain can be computed by Eq. (3.2.19), which is denoted by $\bar{K}_{\mathsf{c}}$; that is, $\bar{K}_{\mathsf{c}} = [189.45 \;\; 5.34]$ with

$$\bar{Q}_{\mathsf{d}} = 10^4 \times \begin{bmatrix} 4.006 & 0.032 \\ 0.032 & 0.001 \end{bmatrix}, \bar{R}_{\mathsf{d}} = 0.2544, \quad \text{and} \quad \bar{V}_{\mathsf{d}} = \begin{bmatrix} -76.863 \\ -1.493 \end{bmatrix}.$$

Fig. 3.10 illustrates the continuous-time states of the plant which uses K_{d} (3.2.8); that is,

$$\begin{aligned} x_{k+1} &= A_f x_k + B_{\mathsf{d}} u_k, \\ y_k &= C x_k, \\ u_k &= K_{\mathsf{d}} \hat{x}_k, \end{aligned}$$

where the specific values of the matrices are the same that of Section 3.2.2. As one can see, while the system trajectories are stabilized to the origin, the fluctuation maintains. On the other hand, when we use the DLQR under consideration of the inter-sample behavior (i.e., K_{c} (3.2.19)), as shown in Fig. (3.11), the fluctuation of the states is disappearing much faster than when the K_{d} is used.

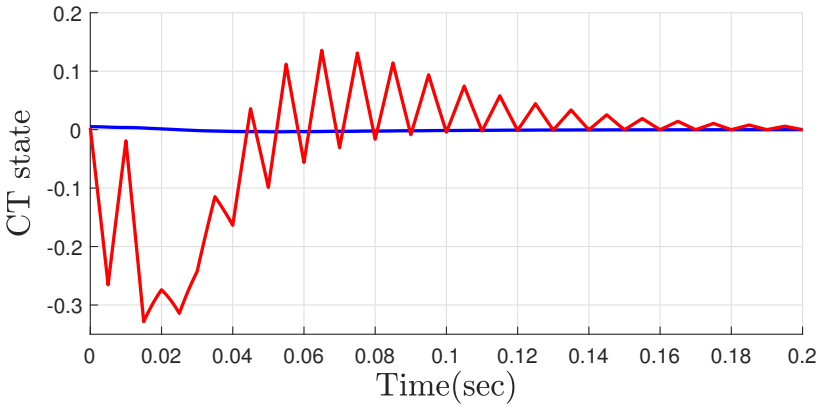


Figure 3.10: Continuous-time state trajectories of the plant when the DLQR with discrete-time performance index is used; that is, $\bar{K}_{\mathsf{d}}$ (3.2.8).

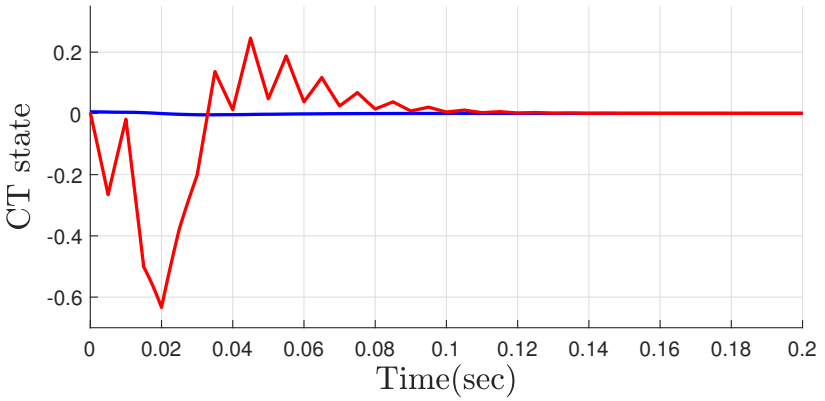


Figure 3.11: Continuous-time state trajectories of the plant when the DLQR with continuous-time performance index is used; that is, $\bar{K}_{\mathsf{c}}$ (3.2.19).

Chapter 4

Masking Attack for Sampled-data System via Input Redundancy

4.1 Problem Formulation

Apart from the system zero-based approaches, in this chapter, we are interested in investigating another type of system property that can be employed to design a fatal cyber-attack when exploited by the attacker. This new property we are concerned with is *input redundancy*, which exists when the target system has a sort of freedom in the input channel compared to the output. Roughly speaking, when the target system has the input redundancy, the way of constructing an attack signal can be separated into two parts; first, an attack signal is selected to enter the system through the redundant inputs and to enforce the system states diverge; next, a secondary signal is added to conceal or mask the influence of the diverging state to the output measurements. For this reason, we call such attack as *masking attack*. It is readily expected that, due to the nature of this attack, the adversary with the masking attack could enjoy the full advantage of other lethal attacks including the zero dynamics attack (i.e., the stealthiness and the ability to disrupt the system).

We consider a continuous-time physical system modeled as

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) + Ed(t), \\ y(t) &= Cx(t) + n(t) \end{aligned} \tag{4.1.1}$$

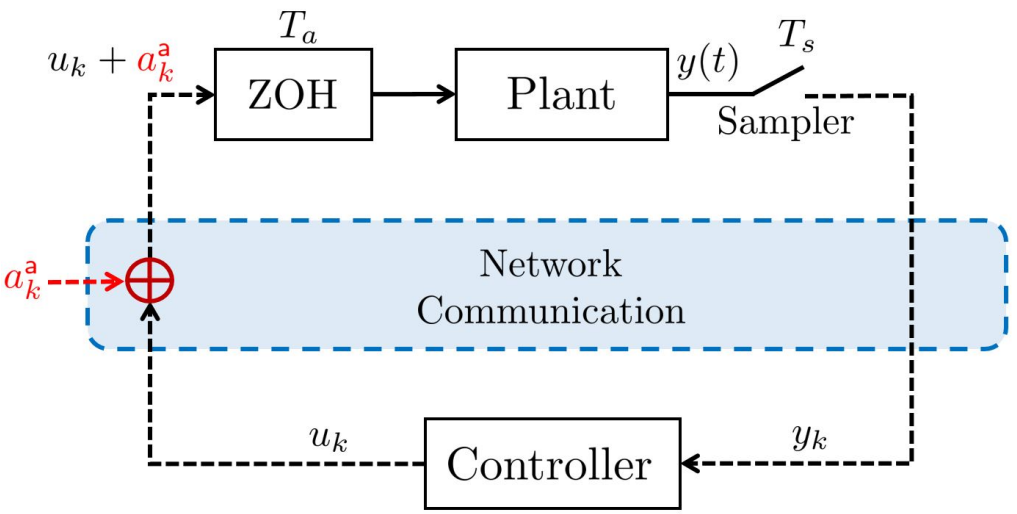


Figure 4.1: Multi-rate sampled-data system connected through a network

where $x \in \mathbb{R}^n$ is the system state, $u \in \mathbb{R}^p$ is the input, $y \in \mathbb{R}^q$ is the output, $d \in \mathbb{R}^r$ is the external disturbance, $n \in \mathbb{R}^q$ is the noise, and A, B, C, and E are constant matrices with appropriate dimensions. It is assumed that the plant (4.1.1) is connected with a discrete-time controller through a communication network that is exposed to a malicious adversary, as seen in Fig. 4.1. It is also assumed that the discrete-time control is performed with the sampler for the output $y(t)$ with the sampling period T_s, and the ZOH for the input $u(t)$ with the sampling period T_a. Note that unlike Chapters 2 and 3, it is allowed the sampling periods of the sampler and ZOH can be different, and this system is called the multi-rate sampled-data system. Therefore, the sampled output is defined

$$y_k = y(kT_s), \quad \forall kT_s \le t < (k+1)T_s, \tag{4.1.2}$$

and the control input $u(t)$ has the form

$$u(t) = u(kT_a) = u_k + a_k^{\mathsf{a}}, \quad \forall kT_a \le t < (k+1)T_a, \tag{4.1.3}$$

where u_k is the output of a discrete-time controller and a_k^{a} is a discrete-time attack signal injected through the vulnerable input communication network. Without

loss of generality, we assume that the attack signal a_k^{a} is initiated at $k = 0$ (or equivalently $t = 0$), while the control system itself may have started operation before $t = 0$. Consequently $x(0)$ can be of arbitrary values.

The problem of our interest is to construct the attack signal a_k^{a} for general multi-rate sampled-data systems where T_s and T_a are not necessarily the same. Specifically, in the dissertation, we allow the ratio between T_s and T_a to a rational number; that is,

$$\mathcal{R} := \frac{T_s}{T_a} \in \mathbb{Q}.$$

In order to deal the ratio $\mathcal{R}$ more easily, throughout this chapter, we often use the coprime fraction $\mathcal{R} = \beta/\alpha$ with $\alpha, \beta \in \mathbb{N}$ (rather than T_s/T_a). Obviously, when $\mathcal{R} \neq 1$, the sensing and the actuation do not occur with the same period. However, even when $\mathcal{R} = 1$, sensing and actuation occurrences may not be synchronized due to the different starting times. To accommodate this case, an *offset* time Δ is introduced in this chapter, which indicates the first sensing time after $t = 0$, and satisfies $0 \leq \Delta < T_s$ (see Fig. 4.2 for illustration). Therefore, the most general case can be effectively described such that, while the actuation time is given by $t = iT_a$, the sensing times are of the form $t = jT_s + \Delta$, $j = 0, 1, \cdots$ so that the control input $u(t)$, attack $a(t)$ and sampled measurement y_j is given by

$$u(t) := u_i, \qquad \forall iT_a \leq t < (i+1)T_a,$$

$$a^{\mathsf{a}}(t) := a_i^{\mathsf{a}}, \qquad \forall iT_a \leq t < (i+1)T_a, \tag{4.1.4}$$

$$y_j := y(jT_s + \Delta). \tag{4.1.5}$$

Throughout this chapter, we usually use i and j to indicate the discrete-times for the ZOH and the sampler, respectively.

For such a class of sampled-data systems, the attack signal a_i^{a} is desired to have the following two properties at once. In order to effectively define these properties, let $\bar{x}(t)$ be the solution of the sampled-data system without any attack (i.e., (4.1.1)–(4.1.5) with $a_i^{\mathsf{a}} \equiv 0$) and let $\bar{y}(t) = C\bar{x}(t)$.

Definition 4.1.1. An attack sequence $\{a_i^{\mathsf{a}}\}_{i=0}^{\infty}$ is said to have *zero-stealthy prop-*

erty if

$$y_j - \bar{y}_j = 0, \quad \forall j \geq 0$$

where $\bar{y}_j := \bar{y}(jT_s + \Delta)$. □

This property directly implies that the plant (4.1.1) under the attack appears to operate normally as if there were no attacks, which means that it is impossible for most conventional anomaly detectors to detect the attack satisfying Definition 4.1.1. This is because as we discussed in Corollary 1.3.1, the underlying principle behind common attack detection schemes is to capture the variation of the residual signal $r_j := C\hat{x}_j - y_j$ where $\hat{x}$ is a state estimate obtained from y and u (so that $C\hat{x}_j$ represents an estimate of the output y_j). Thus, there is no way to recognize the attack injection as long as the attack a_i^{a} ensures $y_j = \bar{y}_j$, since in this case the residual r_j is not altered by the attack signal.

Definition 4.1.2. Let $\{H_k\}_{k=1}^{\infty}$ and $\{\Gamma_k\}_{k=1}^{\infty}$ be sequences of positive numbers H_k and pairwise disjoint sets $\Gamma_k \subset (0, \infty)$, respectively, in which $\bigcup_{k=1}^{\infty} \Gamma_k = (0, \infty)$ and $\sup \Gamma_k = \inf \Gamma_{k+1}$ for all $k = 1, 2, \ldots$. Then, an attack sequence $\{a_i^{\mathsf{a}}\}_{i=0}^{\infty}$ is said to have *disruptive property with* (H_k, Γ_k) if there exists a sequence $\{t_k\}_{k=1}^{\infty}$ of ordered times t_k satisfying $t_k < t_{k+1}$, called *disruption times*, such that $t_k \in \Gamma_k$ and

$$\|x(t_k) - \bar{x}(t_k)\| \geq H_k, \quad \forall k \geq 1.$$

□

The *disruptive property* indicates that the actual state $x(t)$ in continuous-time is forced to be far away from the attack-free one $\bar{x}(t)$ steadily. In the definition, the strength of the attack is characterized by the sequence $\{H_k\}_{k=1}^{\infty}$, whose selection is entirely upon the adversary. On the other hand, the disruption time t_k represents the time when the magnitude of the error variable, $x - \bar{x}$, is larger than H_k. The time interval Γ_k to which t_k belongs determines how frequently the plant state $x(t)$ is perturbed. We note in advance that Γ_k should be carefully selected for the attacker's success, since the attack a_i^{a} is implemented in discrete-time and

thus too many disruption times t_k within one actuation period (or too short time intervals Γ_k) may not be possible. This point will be clarified in Subsection 4.2.1.

At first glance, it may seem that the problem under consideration can be readily tackled by the so-called zero dynamics attack which has been widely studied in the literature [TSSJ12, TSSJ15]. However, this is not true in general, because of the following reasons. First, the zero dynamics attack can be designed only when the target system has an unstable zero, which is possibly violated for a large class of multi-rate sampled-data systems; in fact, it is seen in [NHV15, NHV19] that there may be no unstable zero under the multi-rate operation of the sampler. Moreover, the strength of the zero dynamics attack is determined solely by the location of the unstable zero that is given by the plant's inherent characteristic. Finally, for most of the cases, the zero dynamics attack may not be zero-stealthy in the sense of Definition 4.1.1, because its initiation causes a (small but nonzero) transient that can be observed from the output.

In this chapter, we propose another attack scenario for the sampled-data system that is potentially more lethal than the conventional zero dynamics attack. Our proposal is based on the assumption that the system (4.1.1) with ZOH and sampler has a kind of *input redundancy*. This is the case when the ZOH works faster than the sampler (that is, $\mathcal{R}$ is larger than 1), or the number p of the input channel is larger than that of the output channel, q. Then, the input redundancy makes it possible for the attacker to mask the effect of malicious attack so that the sampled output seems to be normal at each sensing time. As we shall see below, the adversary can generate a masking attack that has the zero-stealthy and disruptive property with arbitrarily large threshold H_k.

4.2 Design of Masking Attack with Zero-stealthy and Disruptive Properties

We begin the attack design by defining a normalized offset $\delta := \Delta/T_s$ (so that $0 \le \delta < 1$), and a new time index (which is a real number) as

$$j_\delta := \delta + \lfloor j - \delta \rfloor, \qquad j = 1, 2, \cdots .$$

Then $j_\delta = (j-1)+\delta$ if $\delta > 0$ and $j_\delta = j$ if $\delta = 0$. Throughout this chapter, we call j_δ above *(shifted) sensing time index*, while i *actuation time index*. Note that using the index j_δ, the sampled-data system can be written as

$$\begin{aligned} x((j_\delta+1)T_s) &= e^{AT_s}x(j_\delta T_s) \\ &\quad + \int_{j_\delta T_s}^{(j_\delta+1)T_s} e^{A((j_\delta+1)T_s-\tau)}(Bu(\tau)+Ed(\tau))d\tau, \\ y(j_\delta T_s) &= Cx(j_\delta T_s) + n(j_\delta T_s) \end{aligned} \tag{4.2.1}$$

for $j = 1, 2, \cdots$, while $x(1_\delta T_s)$ is given by

$$x(1_\delta T_s) = e^{A1_\delta T_s}x(0) + \int_0^{1_\delta T_s} e^{A(1_\delta T_s-\tau)}(Bu(\tau)+Ed(\tau))d\tau.$$

Let us define the error variables as follows:

$$\tilde{x}(t) := x(t) - \bar{x}(t), \quad \tilde{y}(t) := y(t) - \bar{y}(t) = C\tilde{x}(t),$$

in which $\bar{x}$ and $\bar{y}$ are the attack-free state and output as defined in the previous section. The next task for the attack design is to express the $\tilde{x}$-dynamics in two different discrete-time frames. One is associated with the sensing time index j_δ and the other is that with the actuation time index i, which will be used in the attack design and analysis to come. After some computations, one has the error dynamics with j_δ (obtained from (4.2.1)) as

$$\begin{aligned} \tilde{x}((j_\delta+1)T_s) &= e^{AT_s}\tilde{x}(j_\delta T_s) + \int_{j_\delta T_s}^{(j_\delta+1)T_s} e^{A((j_\delta+1)T_s-\tau)}Ba^{\mathtt{a}}(\tau)d\tau, \\ \tilde{y}(j_\delta T_s) &= C\tilde{x}(j_\delta T_s) \end{aligned} \tag{4.2.2}$$

where

$$\tilde{x}(1_\delta T_s) = \int_0^{1_\delta T_s} e^{A(1_\delta T_s-\tau)}Ba^{\mathtt{a}}(\tau)d\tau$$

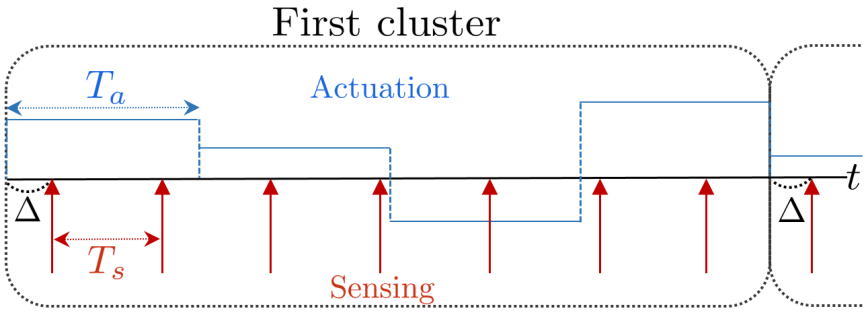


Figure 4.2: Example of a cluster when $\mathcal{R} = T_s/T_a = 4/7 = \beta/\alpha$. There are β actuation times and α sensing times in one cluster.

and the continuous-time signal $a^{\mathsf{a}}(t)$ is defined in (4.1.4). Similarly, we also have

$$\begin{aligned}\tilde{x}(iT_a) &= A_{\mathsf{d}}^{i}\tilde{x}(0) + \sum_{m=0}^{i-1} A_{\mathsf{d}}^{i-1-m} B_{\mathsf{d}} a_m^{\mathsf{a}} \\ &= \sum_{m=0}^{i-1} A_{\mathsf{d}}^{i-1-m} B_{\mathsf{d}} a_m^{\mathsf{a}}\end{aligned} \tag{4.2.3}$$

where the last equality follows from $\tilde{x}(0) = 0$, and

$$A_{\mathsf{d}} := e^{AT_a} \in \mathbb{R}^{n\times n}, \quad B_{\mathsf{d}} := \left(\int_0^{T_a} e^{A\tau} d\tau\right) B \in \mathbb{R}^{n\times p}.$$

Notice that, by the linearity of the plant (4.1.1), the disturbance $d(t)$ and the noise $n(t)$ do not appear in (4.2.2) and (4.2.3). For progression, we need more generalized notations regarding A_{d} and B_{d} which contain the sampling times T_s and T_a both. In this context, let us define the following new notations:

$$A_{\mathsf{d}}^{\langle l,m\rangle} := e^{A(lT_s - mT_a)}, \quad B_{\mathsf{d}}^{\langle l,m\rangle} := \left(\int_0^{lT_s - mT_a} e^{A\tau} d\tau\right) B,$$

where $l, m \in \mathbb{Z}$. From the above definition, A_{d} and B_{d} can also be denoted as $A_{\mathsf{d}}^{\langle 0,-1\rangle}$ and $B_{\mathsf{d}}^{\langle 0,-1\rangle}$, respectively.

4.2.1 Clustering the Time Frame

As aforementioned, for the design of a_i^{a} satisfying Definition 4.1.2, it is important to assign the sequence $\{t_k\}_{k=1}^{\infty}$ of the disruption times in an appropriate way. To this end, let us introduce the concept of 'cluster' in the time framework such that each t_k belongs to each cluster one by one. This is motivated by the observation that as long as the ratio $\mathcal{R}$ is in $\mathbb{Q}$, distribution of the sensing and actuation times always exhibits a certain pattern that repeats in every $\alpha T_s = \beta T_a$ seconds (see Fig. 4.2). Keeping this in mind, the k-th cluster is defined as the time period $(k-1)\beta T_a < t \leq k\beta T_a$ (and sometimes we indicate the left-closed interval $(k-1)\beta T_a \leq t < k\beta T_a$ by calling it the k-th *input*-cluster). It can be seen that each cluster contains exactly α sensing times and β actuation times. As an example, the case of $\mathcal{R} = T_s/T_a = 4/7 = \beta/\alpha$ with non-zero offset $\Delta > 0$ is illustrated in Fig. 4.2.

Throughout this chapter, we set each Γ_k in Definition 4.1.2 as the k-th cluster; namely,

$$\Gamma_k := ((k-1)\beta T_a, k\beta T_a].$$

Then the disruption time t_k associated with Γ_k is given by

$$t_k := (k-1)\beta T_a + T_k^*, \qquad T_k^* \in (0, \beta T_a]$$

(so that $t_k \in \Gamma_k$) where the sequence $\{T_k^*\}_{k=1}^{\infty}$ is chosen by the adversary, and it is often a fixed number, for convenience, such as $T_k^* = \beta T_a$ or $T_k^* = \beta T_a/2$ for all k.

By exploiting the concept of the cluster, we will consider the error dynamics (4.2.2) in terms of the clusters. For this, let us define stacked attacks in the k-th input-cluster, and stacked states and stacked measurements in the k-th cluster, as follows: for $k = 1, 2, \cdots$,

$$a^{\mathsf{a}}\langle k \rangle := \begin{bmatrix} a^{\mathsf{a}}_{(k-1)\beta} \\ a^{\mathsf{a}}_{(k-1)\beta+1} \\ \vdots \\ a^{\mathsf{a}}_{k\beta-1} \end{bmatrix} \in \mathbb{R}^{\beta p},$$

$$\tilde{x}\langle k\rangle := \begin{bmatrix} \tilde{x}(((k-1)\alpha + 1_\delta)T_s) \\ \tilde{x}(((k-1)\alpha + 2_\delta)T_s) \\ \vdots \\ \tilde{x}(((k-1)\alpha + \alpha_\delta)T_s) \end{bmatrix} \in \mathbb{R}^{\alpha n},$$
$$\tilde{y}\langle k\rangle := \mathcal{C}\tilde{x}\langle k\rangle \in \mathbb{R}^{\alpha q}$$

where $\mathcal{C} := I_\alpha \otimes C$ (I_α is the identity matrix of size α and $\otimes$ is the Kronecker product). It is noted that the vector $\tilde{y}\langle k\rangle$ is the collection of α measurements within one cluster.

Now, let us focus on the terminal state of each cluster, which is denoted by

$$\tilde{x}_k^{\mathsf{c}} := \tilde{x}(k\beta T_a) \in \mathbb{R}^n.$$

Then, from (4.2.3), one can derive that

$$\begin{aligned} \tilde{x}_k^{\mathsf{c}} &= A_{\mathsf{d}}^{\beta}\tilde{x}_{k-1}^{\mathsf{c}} + \begin{bmatrix} A_{\mathsf{d}}^{\beta-1}B_{\mathsf{d}} & A_{\mathsf{d}}^{\beta-2}B_{\mathsf{d}} & \cdots & B_{\mathsf{d}} \end{bmatrix} a^{\mathsf{a}}\langle k\rangle \\ &=: A_{\mathsf{d}}^{\beta}\tilde{x}_{k-1}^{\mathsf{c}} + \Phi_{\mathsf{c}}a^{\mathsf{a}}\langle k\rangle \end{aligned} \tag{4.2.4}$$

with $\tilde{x}_0^{\mathsf{c}} = 0$. Similarly, one can derive the following for $\tilde{x}\langle k\rangle$.

Lemma 4.2.1. It follows that

$$\tilde{x}\langle k\rangle = \bar{A}_\alpha\tilde{x}_{k-1}^{\mathsf{c}} + \Pi a^{\mathsf{a}}\langle k\rangle \tag{4.2.5}$$
$$\tilde{y}\langle k\rangle = \mathcal{C}\bar{A}_\alpha\tilde{x}_{k-1}^{\mathsf{c}} + \mathcal{C}\Pi a^{\mathsf{a}}\langle k\rangle \tag{4.2.6}$$

where

$$\bar{A}_\alpha := \begin{bmatrix} e^{A1_\delta T_s} \\ e^{A2_\delta T_s} \\ \vdots \\ e^{A\alpha_\delta T_s} \end{bmatrix}$$

and the (l,m)-th block of $\Pi \in \mathbb{R}^{\alpha n \times \beta p}$ is defined as

$$\Pi(l,m) := \begin{cases} A_{\mathsf{d}}^{\langle l_\delta, m\rangle} B_{\mathsf{d}}, & m = 1, \cdots, \lfloor l_\delta \mathcal{R} \rfloor, \\ B_{\mathsf{d}}^{\langle l_\delta, \lfloor l_\delta \mathcal{R} \rfloor \rangle}, & m = \lfloor l_\delta \mathcal{R} \rfloor + 1, \\ 0, & m = \lfloor l_\delta \mathcal{R} \rfloor + 2, \cdots, \beta, \end{cases} \tag{4.2.7}$$

for $l = 1, \cdots, \alpha$. □

Proof: Consider the first cluster $k = 1$, in which the state $\tilde{x}(0)$ at the beginning of the cluster is zero and $\tilde{x}_0^{\mathsf{c}} = 0$. With the property $\lfloor j_\delta \mathcal{R} \rfloor T_a \le j_\delta \mathcal{R} T_a = j_\delta T_s$, one can compute the state $\tilde{x}(j_\delta T_s)$, whose sensing time $j_\delta T_s$ belongs to this cluster, by the variation of constant formula as follows:

$$\begin{aligned} \tilde{x}(j_\delta T_s) &= \int_0^{T_a} e^{A(j_\delta T_s - \tau)} d\tau B a_0^{\mathsf{a}} + \int_{T_a}^{2T_a} e^{A(j_\delta T_s - \tau)} d\tau B a_1^{\mathsf{a}} \\ &\quad + \cdots + \int_{(\lfloor j_\delta \mathcal{R} \rfloor - 1)T_a}^{(\lfloor j_\delta \mathcal{R} \rfloor) T_a} e^{A(j_\delta T_s - \tau)} d\tau B a_{(\lfloor j_\delta \mathcal{R} \rfloor - 1)}^{\mathsf{a}} \\ &\quad + \int_{\lfloor j_\delta \mathcal{R} \rfloor T_a}^{j_\delta T_s} e^{A(j_\delta T_s - \tau)} d\tau B a_{\lfloor j_\delta \mathcal{R} \rfloor}^{\mathsf{a}} \\ &= \sum_{m=1}^{\lfloor j_\delta \mathcal{R} \rfloor} e^{A(j_\delta T_s - m T_a)} \int_{(m-1)T_a}^{m T_a} e^{A(m T_a - \tau)} d\tau B a_{(m-1)}^{\mathsf{a}} \\ &\quad + \int_0^{j_\delta T_s - \lfloor j_\delta \mathcal{R} \rfloor T_a} e^{A\tau} d\tau B a_{\lfloor j_\delta \mathcal{R} \rfloor}^{\mathsf{a}} \\ &= \sum_{m=1}^{\lfloor j_\delta \mathcal{R} \rfloor} A_{\mathsf{d}}^{\langle j_\delta, m \rangle} B_{\mathsf{d}} a_{(m-1)}^{\mathsf{a}} + B_{\mathsf{d}}^{\langle j_\delta, \lfloor j_\delta \mathcal{R} \rfloor \rangle} a_{\lfloor j_\delta \mathcal{R} \rfloor}^{\mathsf{a}}. \end{aligned} \tag{4.2.8}$$

When $\lfloor 1_\delta \mathcal{R} \rfloor = 0$ (which happens if $j = 1$ and $1_\delta T_s < T_a$), it should be interpreted that the summation term in the above equation is zero or null. The discussion so far verifies (4.2.5) and the matrix Π for $k = 1$.

For the general k-th clusters ($k > 1$), the derivation is the same (because the pattern for actuation and sensing times are repeated along the clusters) except that the state $\tilde{x}_{k-1}^{\mathsf{c}} = \tilde{x}((k-1)\beta T_a)$ need not be zero. Taking into account $\tilde{x}_{k-1}^{\mathsf{c}}$ as the initial condition for the corresponding cluster, one can easily verify (4.2.5)

for $k > 1$. Once (4.2.5) is verified, (4.2.6) trivially follows. ■

Now for simplicity of presentation, let us normalize the disruption time as $t_k^* := T_k^*/(\beta T_a) \in (0, 1]$. Then, the error state at the disruption time t_k, which we will denote as $\tilde{x}_k^{\mathsf{a}} := \tilde{x}(t_k)$, is computed as follows.

Lemma 4.2.2. It follows that

$$\tilde{x}_k^{\mathsf{a}} = \bar{A}_k^* \tilde{x}_{k-1}^{\mathsf{c}} + \Phi_k^* a^{\mathsf{a}} \langle k \rangle \tag{4.2.9}$$

where $\bar{A}_k^* := e^{AT_k^*} = e^{At_k^* \beta T_a}$ and

$$\Phi_k^* = [\Phi_k^*(1,1), \cdots, \Phi_k^*(1,\beta)] \in \mathbb{R}^{n \times \beta p} \tag{4.2.10}$$

$$\text{where } \Phi_k^*(1,m) = \begin{cases} A_{\mathsf{d}}^{\langle 0, m-\beta t_k^* \rangle} B_{\mathsf{d}}, & m = 1, \cdots, \lfloor \beta t_k^* \rfloor \\ B_{\mathsf{d}}^{\langle 0, \lfloor \beta t_k^* \rfloor - \beta t_k^* \rangle}, & m = \lfloor \beta t_k^* \rfloor + 1 \\ 0, & m = \lfloor \beta t_k^* \rfloor + 2, \cdots, \beta. \end{cases}$$

□

Proof: The proof is similarly done as Lemma 4.2.1. For the first cluster ($k = 1$), the state $\tilde{x}_1^{\mathsf{a}}$ (i.e., error state at time t_1) is evaluated similarly as (4.2.8) with $j_\delta T_s$ being replaced by $t_1 = \beta t_1^* T_a$, and $j_\delta \mathcal{R}$ being replaced by βt_1^*. Indeed, it follows that

$$\begin{aligned} \tilde{x}_1^{\mathsf{a}} &= \tilde{x}(t_1) = \tilde{x}(T_1^*) \\ &= \sum_{m=1}^{\lfloor \beta t_1^* \rfloor} e^{A(\beta t_1^* T_a - m T_a)} \int_{(m-1)T_a}^{mT_a} e^{A(mT_a - \tau)} d\tau B a_{m-1}^{\mathsf{a}} \\ &\quad + \int_0^{\beta t_1^* T_a - \lfloor \beta t_1^* \rfloor T_a} e^{A\tau} d\tau B a_{\lfloor \lfloor \beta t_1^* \rfloor \rfloor}^{\mathsf{a}} \\ &= \sum_{m=1}^{\lfloor \beta t_1^* \rfloor} A_{\mathsf{d}}^{\langle 0, m-\beta t_1^* \rangle} B_{\mathsf{d}} a_{m-1}^{\mathsf{a}} + B_{\mathsf{d}}^{\langle 0, \lfloor \beta t_1^* \rfloor - \beta t_1^* \rangle} a_{\lfloor \beta t_1^* \rfloor}^{\mathsf{a}}. \end{aligned}$$

Similar in Lemma 4.2.1, if $\lfloor \beta t_1^* \rfloor = 0$, the summation term in the above equation becomes zero. Thus, (4.2.9) and (4.2.10) are verified for the first cluster.

For the case $k > 1$, by taking into account the initial condition $\tilde{x}_{k-1}^{\mathsf{c}}$ and by

noting that the matrix Φ_k^* is obtained exactly the same way as for $k = 1$, equation (4.2.9) is easily verified. ■

Note that, if all T_k^* are chosen as a constant for all $k \geq 1$, then both $\bar{A}_k^*$ and Φ_k^* are constant matrices. Now, with Lemma 4.2.1 and Lemma 4.2.2, the problem of our interest is reformulated in the cluster-wise sense; i.e., our interest becomes designing an attack sequence $a^{\mathsf{a}}\langle k\rangle$ that satisfies $\|\tilde{x}_k^{\mathsf{a}}\| \geq H_k$ (disruptive property), and at the same time, $\tilde{y}\langle k\rangle \equiv 0$ for each k-th cluster (zero-stealthy property) for all $k \geq 1$.

Remark 4.2.1. In fact, the concept of clustering is required mainly to extend the class of sampled-data systems to those having $\mathcal{R} \in \mathbb{Q}$ and $\Delta \neq 0$, which is one of the main contributions of the dissertation. When it comes to $\mathcal{R} \in \mathbb{N}$ (so that $\alpha = 1$) and $\Delta = 0$ as in the conference paper [KPSE16], clustering the time frame is not necessary for attack design and a large part of the notations used in this section can be simplified (e.g., in that case, the k-th cluster turns out to be the k-th sampling period $(k-1)T_s < t \leq kT_s$). □

4.2.2 Conditions for Masking Attack Design

With equations (4.2.5), (4.2.6), and (4.2.9) at hand, conditions for attack design can be established. First of all, by (4.2.6), stealthiness of the attack is obtained if the attack sequence $a^{\mathsf{a}}\langle k\rangle$ for the k-th cluster belongs to the kernel of $\mathcal{C}\Pi$, and so, we require the kernel is non-trivial. Second, for the disruptive property of the state $\tilde{x}_k^{\mathsf{a}}$ in (4.2.9), we ask the kernel of Φ_k^* not to include the kernel of $\mathcal{C}\Pi$ because, if $\ker \Phi_k^* \supset \ker \mathcal{C}\Pi$, then any stealthy attack has no affect on $\tilde{x}_k^{\mathsf{a}}$. Finally, as the attack is initiated, the state $\tilde{x}(t)$ becomes non-zero, and therefore, even if the attack $a^{\mathsf{a}}\langle k-1\rangle$ is designed to be stealthy from the measurement vector $\tilde{y}\langle k-1\rangle$ for the $(k-1)$-th cluster, it may become detectable through non-zero $\tilde{x}_{k-1}^{\mathsf{c}} = \tilde{x}((k-1)\beta T_a)$ in the k-th cluster. See (4.2.5) and (4.2.6). In order to counteract it, we require the range space of $\mathcal{C}\bar{A}_\alpha$ would belong to the range space of $\mathcal{C}\Pi$ so that some component of the attack sequence $a^{\mathsf{a}}\langle k\rangle$ is designed to cancel the effect of x_{k-1}^{c} on $\tilde{y}\langle k\rangle$. These discussions yield the following formal assumption.

Assumption 4.2.1. The following conditions hold:

(a) $\ker \mathcal{C}\Pi \neq \{0\}$,

(b) $\ker \mathcal{C}\Pi \not\subset \ker \Phi_k^*$, $\forall k \geq 1$, with a sequence $\{t_k^*\}_{k=1}^{\infty}$ of normalized disruption times,

(c) $\mathrm{im}\, \mathcal{C}\bar{A}_\alpha \subset \mathrm{im}\, \mathcal{C}\Pi$.

□

Although at first glance the conditions for Assumption 4.2.1 might seem hard to check, it can be easily verified by a few sufficient conditions that are derived in the following. For the item (a), $\alpha q < \beta p$ (or, $q < \mathcal{R}p$) implies $\ker \mathcal{C}\Pi \neq \{0\}$ because $\mathcal{C}\Pi \in \mathbb{R}^{\alpha q \times \beta p}$ becomes a fat matrix. It is interesting to see that item (a) signifies the *input redundancy*, since it is satisfied either when the number p of inputs is large, or when the actuator works faster than the sensor (i.e., $\mathcal{R} = T_s/T_a = \beta/\alpha$ is large enough). Hence, a sufficient condition for the item (a) is obviously $qT_a < pT_s$, which is simpler to check than item (a). On the other hand, it is noted that the condition (c) holds if the matrix $\mathcal{C}\Pi$ has full row rank or if the matrix Π has full row rank. Finally, for the condition (b), we have the following.

Proposition 4.2.3. If the condition (a) of Assumption 4.2.1 holds and B_{d} has full column rank (i.e., rank $B_{\mathsf{d}} = p$), then there exists a sequence $\{t_k^*\}_{k=1}^{\infty}$ with which the condition (b) holds. □

Proof: By the condition (a), pick any non-zero $z = \mathrm{col}(z_1, \cdots, z_\beta) \in \ker \mathcal{C}\Pi$ where $z_i \in \mathbb{R}^p$. Define the index $i^* := \min\{i : z_i \neq 0, i = 1, \ldots, \beta\}$, and pick the disruption time $t_k^* \in (0, 1]$ such that $i^* = \beta t_k^*$. Then, it follows from (4.2.10) that $\Phi_k^* z = B_{\mathsf{d}} z_{i^*} \neq 0$ since B_{d} has full column rank. This implies that $\ker \mathcal{C}\Pi \not\subset \ker \Phi_k^*$, i.e., the item (b). ■

From the discussions so far, a sufficient condition for Assumption 4.2.1 can be presented as follows:

Corollary 4.2.4. With appropriate normalized disruption times $\{t_k^*\}_{k=1}^{\infty}$, Assumption 4.2.1 is satisfied if the following three conditions hold:

(a)* $qT_a < pT_s$;

(b)* rank $B_{\mathsf{d}} = p$;

(c)* $\mathcal{C}\Pi$ or Π has full row rank.

□

Remark 4.2.2. As a special case, let us consider the case when $\mathcal{R}$ is a positive integer (i.e., $\mathcal{R} = N \geq 1$ so that $\alpha = 1$ and $\beta = N$), and $\delta = 0$. In this case, the notations can be simplified as follows:

$$\begin{aligned}
\mathcal{C} &= C \\
\bar{A}_\alpha &= e^{AT_s} = A_{\mathsf{d}}^N \\
\Pi &= [e^{A(T_s - T_a)} B_{\mathsf{d}}, e^{A(T_s - 2T_a)} B_{\mathsf{d}}, \cdots, e^{A(T_s - (N-1)T_a)} B_{\mathsf{d}}, B_{\mathsf{d}}] \\
&= [A_{\mathsf{d}}^{N-1} B_{\mathsf{d}}, A_{\mathsf{d}}^{N-2} B_{\mathsf{d}} \cdots, B_{\mathsf{d}}].
\end{aligned}$$

Then, the conditions (a) and (c) of Assumption 4.2.1 can be read as

(a) $\{0\} \neq \ker C[A_{\mathsf{d}}^{N-1} B_{\mathsf{d}}, \cdots, B_{\mathsf{d}}]$,

(c) $\text{im } CA_{\mathsf{d}}^N \subset \text{im } C[A_{\mathsf{d}}^{N-1} B_{\mathsf{d}}, \cdots, B_{\mathsf{d}}]$.

It is clear that the above conditions hold if $q < Np$ and if either $C[A_{\mathsf{d}}^{N-1} B_{\mathsf{d}}, \cdots, B_{\mathsf{d}}]$ or $[A_{\mathsf{d}}^{N-1} B_{\mathsf{d}}, \cdots, B_{\mathsf{d}}]$ has full row rank. Furthermore, the disruption time t_k^* (in the assumption) can be chosen as one of $\{1/N, 2/N, \cdots, 1\}$. By choosing such t_k^* (i.e., $t_k^* \in \{j/N |\ j = 1, \cdots, N\}$), the matrix Φ_k^* (in (4.2.10)) becomes

$$\Phi_k^*|_{t_k^* = \frac{j}{N}} = [A_{\mathsf{d}}^{j-1} B_{\mathsf{d}}, A_{\mathsf{d}}^{j-2} B_{\mathsf{d}}, \cdots, B_{\mathsf{d}}, 0, \cdots, 0].$$

To facilitate selection of t_k^* among the candidates, we consider the condition given by

$$\text{(b')} \qquad \ker C[A_{\mathsf{d}}^{N-1} B_{\mathsf{d}}, \cdots, B_{\mathsf{d}}] \quad \not\subset \quad \ker \begin{bmatrix} B_{\mathsf{d}} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_{\mathsf{d}}^{N-2} B_{\mathsf{d}} & A_{\mathsf{d}}^{N-3} B_{\mathsf{d}} & \cdots & 0 \\ A_{\mathsf{d}}^{N-1} B_{\mathsf{d}} & A_{\mathsf{d}}^{N-2} B_{\mathsf{d}} & \cdots & B_{\mathsf{d}} \end{bmatrix}.$$

When this condition holds, one can pick suitable t_k^* among the candidates for the condition (b) of Assumption 4.2.1. Another sufficient condition for (b) in this special case is: (b") $\ker C \cap \text{ im } \Pi \neq \{0\}$. This is because (b") means that there exists a vector v such that $\Pi v \neq 0$ and $\Pi v \in \ker C$. This implies that the vector v belongs to $\ker C\Pi$ while it does not belong to $\ker \Pi$, which guarantees (b') with $t_k^* = 1$. In Section 4.3.1, we demonstrate this case with $N = 1$. □

4.2.3 Off-line Construction of Attack Signal

In this subsection, based on Assumption 4.2.1, we design an attack sequence a_i^{a}, or equivalently $a^{\mathsf{a}}\langle k\rangle$, that solves the reformulated problem; i.e., to guarantee $\|x_k^{\mathsf{a}}\| \geq H_k$ and $\tilde{y}\langle k\rangle \equiv 0$ for $k = 1, 2, \cdots$. In particular, we propose the sequence $a^{\mathsf{a}}\langle k\rangle$ in the following form:

$$a^{\mathsf{a}}\langle k\rangle = \kappa_k \eta_{\langle k\rangle} + \zeta_{\langle k\rangle} \in \mathbb{R}^{\beta p} \tag{4.2.11}$$

where κ_k is a positive constant and $\eta_{\langle k\rangle}, \zeta_{\langle k\rangle} \in \mathbb{R}^{\beta p}$. Intuitively speaking, κ_k and $\eta_{\langle k\rangle}$ are used to disrupt the system, while $\zeta_{\langle k\rangle}$ *masks* the effect of $\kappa_k \eta_{\langle k\rangle}$ for the stealthiness. The specific idea is to pick $\eta_{\langle k\rangle}$ such that $\Pi\eta_{\langle k\rangle}$ is stealthy (i.e., belongs to $\ker \mathcal{C}$) but disruptive (i.e., $\Phi_k^* \eta_{\langle k\rangle} \neq 0$) while κ_k decides the intensity of disruption, and to pick $\zeta_{\langle k\rangle}$ to counteract the effect of non-zero $\tilde{x}_{k-1}^{\mathsf{c}}$ on $\tilde{y}\langle k\rangle$ (i.e., $\bar{A}_\alpha \tilde{x}_{k-1}^{\mathsf{c}} + \Pi\zeta_{\langle k\rangle} \in \ker \mathcal{C}$). The geometric meaning of each component of $a^{\mathsf{a}}\langle k\rangle$ is illustrated in Fig. 4.3.

At this stage, it should be noticed that the attacker requires the exact values of the terminal state $\tilde{x}_k^{\mathsf{c}}$ of each cluster in order to determine $\zeta_{\langle k+1\rangle}$. At first glance, computing $\tilde{x}_k^{\mathsf{c}} = \tilde{x}(k\beta T_a)$ off-line looks impossible because $\tilde{x}(t)$ is defined as the difference between the actual state $x(t)$ and the (attack-free) virtual one $\bar{x}(t)$. Yet interestingly, one can always compute $\tilde{x}_k^{\mathsf{c}}$ only from the pre-determined attack signals $a^{\mathsf{a}}\langle 1\rangle$, ..., $a^{\mathsf{a}}\langle k\rangle$, without measuring $x(t)$ and $\bar{x}(t)$ directly. This is because the dynamics (4.2.4) that generates the terminal state $\tilde{x}_k^{\mathsf{c}}$ has the zero initial condition $\tilde{x}_0^{\mathsf{c}} = 0$ (by definition), and the adversary already has the inputs $a^{\mathsf{a}}\langle 1\rangle, \ldots, a^{\mathsf{a}}\langle k\rangle$, and thus it is easy to compute $\tilde{x}_k^{\mathsf{c}}$ by duplicating the dynamics (4.2.4). Keeping this in mind, in what follows we explicitly utilize $\tilde{x}_k^{\mathsf{c}}$ in the attack

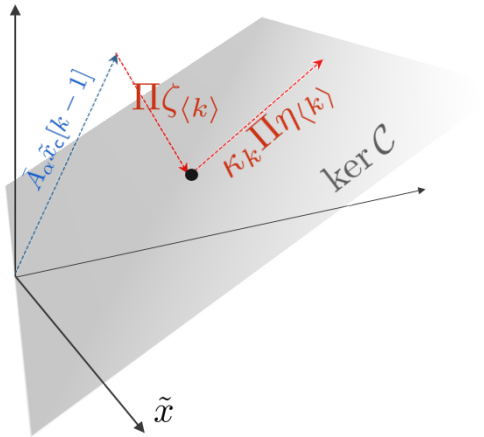


Figure 4.3: Graphical interpretation of attack components.

design.

The attack signal is designed sequentially, i.e., in the order of $a^{\mathsf{a}}\langle 1\rangle$, $a^{\mathsf{a}}\langle 2\rangle$, and so on. As the first step, let $\zeta_{\langle 1\rangle} = 0$ (since there is no attack before the time $t = 0$), and pick $\eta_{\langle 1\rangle} \in \ker \mathcal{C}\Pi$ such that $\Phi_1^* \eta_{\langle 1\rangle} \neq 0$ (whose existence is guaranteed by Assumption 4.2.1.(a)). Then, the stealthiness follows since

$$\tilde{y}\langle 1\rangle = \mathcal{C}\Pi a^{\mathsf{a}}\langle 1\rangle = \kappa_1 \mathcal{C}\Pi\eta_{\langle 1\rangle} = 0.$$

Subsequently, for the disruptive property, pick $\kappa_1 > 0$ such that

$$\|\tilde{x}_1^{\mathsf{a}}\| = \kappa_1 \|\Phi_1^* \eta_{\langle 1\rangle}\| \geq H_1. \tag{4.2.12}$$

By doing this, the attack signal, $a^{\mathsf{a}}\langle 1\rangle = \mathrm{col}(a_0^{\mathsf{a}}; \cdots; a_{\beta-1}^{\mathsf{a}})$, having the stealthy and disruptive properties is obtained for the first cluster $(0, \beta T_a]$.

Next, in order to design $a^{\mathsf{a}}\langle 2\rangle$ in the second cluster, we consider the second stacked output $\tilde{y}\langle 2\rangle$, which is given by

$$\tilde{y}\langle 2\rangle = \mathcal{C}\bar{A}_\alpha \tilde{x}_1^{\mathsf{c}} + \mathcal{C}\Pi a^{\mathsf{a}}\langle 2\rangle \tag{4.2.13}$$

where $\tilde{x}_1^{\mathsf{c}}$ is computed by (4.2.4) and $a^{\mathsf{a}}\langle 2\rangle = \kappa_2 \eta_{\langle 2\rangle} + \zeta_{\langle 2\rangle}$ (see (4.2.11)). Similar

to before, we pick $\eta_{\langle 2\rangle}$ such that

$$\mathcal{C}\Pi\eta_{\langle 2\rangle} = 0 \ \text{ and } \ \Phi_2^*\eta_{\langle 2\rangle} \neq 0.$$

Subsequently, in order to conceal the residual effect of $\tilde{x}_1^{\mathsf{c}}$ in (4.2.13), pick $\zeta_{\langle 2\rangle}$ satisfying

$$\mathcal{C}\Pi\zeta_{\langle 2\rangle} = -\mathcal{C}\bar{A}_\alpha\tilde{x}_1^{\mathsf{c}}. \tag{4.2.14}$$

Now, the error state in the second cluster is computed as

$$\begin{aligned}\tilde{x}_2^{\mathsf{a}} &= \bar{A}_2^*\tilde{x}_1^{\mathsf{c}} + \Phi_2^* a^{\mathsf{a}}\langle 2\rangle \\ &= \bar{A}_2^*\tilde{x}_1^{\mathsf{c}} + \Phi_2^*\zeta_{\langle 2\rangle} + \Phi_2^*\kappa_2\eta_{\langle 2\rangle},\end{aligned}$$

and to achieve the disruptive property, we take κ_2 such that

$$\kappa_2\|\Phi_2^*\eta_{\langle 2\rangle}\| \geq H_2 + \|\bar{A}_2^*\tilde{x}_1^{\mathsf{c}} + \Phi_2^*\zeta_{\langle 2\rangle}\|. \tag{4.2.15}$$

This ensures the zero-stealthy and disruptive properties of the attack in the second cluster $(\beta T_a, 2\beta T_a]$.

We now generalize the procedure.

- **Procedure of Attack Signal Generation**:
  Step k $(k = 1, 2\ldots)$: Take $\zeta_{\langle k\rangle}$ satisfying the following equation:

$$\mathcal{C}\Pi\zeta_{\langle k\rangle} = -\mathcal{C}\bar{A}_\alpha\tilde{x}_{k-1}^{\mathsf{c}} \tag{4.2.16}$$

  (for $k = 1$, $\tilde{x}_0^{\mathsf{c}} = 0$ so that $\zeta_{\langle 1\rangle} = 0$).
  Pick $\eta_{\langle k\rangle}$ satisfying both

$$\eta_{\langle k\rangle} \in \ker\mathcal{C}\Pi \ \text{ and } \ \eta_{\langle k\rangle} \notin \ker\Phi_k^*,$$

  and then select a positive κ_k such that

$$\kappa_k \geq \frac{H_k + \|\bar{A}_k^*\tilde{x}_{k-1}^{\mathsf{c}} + \Phi_k^*\zeta_{\langle k\rangle}\|}{\|\Phi_k^*\eta_{\langle k\rangle}\|}. \tag{4.2.17}$$

With these terms, construct $a^{\mathsf{a}}\langle k\rangle = \kappa_k \eta_{\langle k\rangle} + \zeta_{\langle k\rangle}$ and update

$$\tilde{x}_k^{\mathsf{c}} = A_{\mathsf{d}}^{\beta} \tilde{x}_{k-1}^{\mathsf{c}} + \Phi_{\mathsf{c}} a^{\mathsf{a}}\langle k\rangle$$

by (4.2.4). □

Remark 4.2.3. It is noted that the construction of an attack sequence can be done off-line, or *a priori* before the attack begins because the procedure does not need any real-time information. Moreover, if the normalized disruption time $t_k^* \in (0, 1]$ is chosen as a fixed constant for all $k \geq 1$, then the matrices Φ_k^* are the same for all $k \geq 1$. Then, the vector $\eta_{\langle k\rangle}$ can also be chosen as a constant η. □

We close this section by summarizing the discussions so far.

Theorem 4.2.5. Suppose that the adversary has the information of T_s, T_a, and Δ as well as the system information of A, B, and C. Assume also that $\mathcal{R} = T_s/T_a \in \mathbb{Q}$ and Assumption 4.2.1 holds with a sequence $\{t_k^*\}_{k=1}^{\infty}$ of normalized disruption times $t_k^* \in (0, 1]$. Then for $\Gamma_k = ((k-1)\beta T_a, k\beta T_a]$ and any given $\{H_k\}_{k=1}^{\infty}$, the attack sequence $\{a_i^{\mathsf{a}}\}_{i=0}^{\infty}$ constructed via the proposed procedure has the zero-stealthy property and the disruptive property with (H_k, Γ_k). □

Proof: The proof is replaced by the design procedure of attack signal. ■

Remark 4.2.4. While the basic idea of masking attack was briefly mentioned in [NHV15], this dissertation presents a further development of the idea of [NHV15] in the following two senses. Firstly, the input redundancy needed in the attack design is relaxed (i.e., the condition (a) of Assumption 4.2.1). That is, while the work of [NHV15] focused only on the number of inputs and outputs, in this work the ratio between T_s and T_a also comes into the picture. As a result, the requirement $p > q$ for the attack design in [NHV15] can be relaxed into $pT_s > qT_a$ (i.e., the condition (a)* in Corollary 4.2.4), and this new condition is achieved when $\mathcal{R} = \beta/\alpha = T_s/T_a$ is sufficiently large (even if $p > q$ is not met). Secondly, we present a recursive design method of the masking attack having zero-stealthy and disruptive property. □

4.2.4 Practical Stealthiness of Masking Attack with $\mathcal{R} \in \mathbb{R}$

Occasionally, the ratio of actuation to sampling period, $\mathcal{R}$, can be a *real number* (i.e., $\mathcal{R} = T_s/T_a \in \mathbb{R}$). The stealthiness of the proposed masking attack is based on the concept of clustering the time frame in which the periodical encountering time of actuation and sensing is considered, i.e., $\alpha T_s = \beta T_a$. In the case of $\mathcal{R} \in \mathbb{R}$, however, the presented masking attack cannot achieve the zero-stealthy property Definition 4.1.1 since the periodical encountering time does not exist. Instead, by approximating the ratio $\mathcal{R}$ with a truncation function, the proposed attack $\{a_i^{\mathsf{a}}\}_{i=0}^{\infty}$ can be constructed, which has zero-stealthy property in a practical sense until a certain time. To formalize this notion, we define a practical stealthiness as follows.

Definition 4.2.1. For given threshold ϵ, an attack sequence $\{a_i^{\mathsf{a}}\}_{i=0}^{\infty}$ is said to have *practical-stealthy property* with $t_{\text{sth}} > 0$ if $\|\tilde{y}(j_\delta T_s)\|_\infty < \epsilon$ for $j_\delta T_s \leq t_{\text{sth}}$. Also, the time t_{sth} is called *stealthy assurance time.* □

In what follows, we consider how long the designed attack can be maintained in stealthy. Computing the difference of the approximated and actual outputs, we compute the stealthy assurance time t_{sth} when the threshold ϵ is given and the employed attack is bounded to a specific value, i.e., $\|a_i^{\mathsf{a}}\| \leq \bar{a}^{\mathsf{a}}$. In order to achieve approximate model, consider a truncation function:

$$\text{trunc}(\mathcal{R}, m) := \frac{\lfloor 10^m \cdot \mathcal{R} \rfloor}{10^m} =: \tilde{\mathcal{R}} = \frac{\tilde{\beta}}{\tilde{\alpha}} \in \mathbb{Q} \tag{4.2.18}$$

where $\mathcal{R} \in \mathbb{R}$, $\tilde{\beta}/\tilde{\alpha}$ is coprime fraction with $\tilde{\alpha}, \tilde{\beta} \in \mathbb{N}$. This truncation approximates the ratio $\mathcal{R} = T_s/T_a \in \mathbb{R}$ into a rational number. The first task after the truncation is to determine the actuation and sensing periods corresponding to $\tilde{\mathcal{R}}$. We fix the actuation period T_a and recompute the new sensing period $\tilde{T}_s$ such that $\tilde{T}_s := T_a\tilde{\mathcal{R}} \in \mathbb{R}$. By doing so, we can achieve rational ratio between actuation and sensing times. This is useful for attack designing in that distribution of the actuation and sensing times has a pattern that repeats in every $\tilde{\alpha}\tilde{T}_s = \tilde{\beta}T_a$ seconds. In addition, in accordance with $\tilde{T}_s$, the offset δ need to be redefined by

$\tilde{\delta}$ as follows:

$$\tilde{\delta} := \tilde{\Delta}/\tilde{T}_s \quad \text{where} \quad \tilde{\Delta} := \begin{cases} \Delta, & \text{if } \Delta < \tilde{T}_s \\ \Delta - \tilde{T}_s, & \text{if } \Delta \geq \tilde{T}_s \end{cases}$$

so that $j_{\tilde{\delta}} := \tilde{\delta} + \lfloor j - \tilde{\delta} \rfloor$. Now again, using the approximate sensing period $\tilde{T}_s$ and actuation period T_a, the masking attack can be designed by following the attack generation procedure in Subsection 4.2.3. Then, the resultant attack sequence is zero-stealthy at each sensing time $j_{\tilde{\delta}}\tilde{T}_s$ (i.e., $y(j_{\tilde{\delta}}\tilde{T}_s) \equiv 0$), and thus, by computing the output deviation at each sensing time which is caused by approximation of the ratio, the stealthy assurance time can be achieved.

Lemma 4.2.6. Suppose the adversary knows the threshold of the anomaly detector, $\epsilon > 0$. Then, for a given truncation level m and input bound $\bar{a}^{\mathsf{a}}$, the stealthy insurance time t_{sth} is given by

$$t_{\text{sth}} = \arg\max_j \left(\left\| C(e^{Aj_\delta(\mathcal{R}-\tilde{\mathcal{R}})T_a} - I) \right\|_M \mathcal{A}_j + \int_{j_\delta \tilde{\mathcal{R}} T_a}^{j_\delta \mathcal{R} T_a} \left\| C e^{A(j_\delta T_s - \tau)} B \right\|_M d\tau p \bar{a}^{\mathsf{a}} < \epsilon \right) \tag{4.2.19}$$

where

$$\mathcal{A}_j := \int_0^{j_\delta T_s} \|e^{A\tau} B\|_1 d\tau p \bar{a}^{\mathsf{a}}.$$

□

Proof: The error state at actual sensing time $j_\delta T_s$ is derived as

$$\tilde{x}(j_\delta T_s) = e^{A(j_\delta T_s - j_{\tilde{\delta}}\tilde{T}_s)} x(j_{\tilde{\delta}}\tilde{T}_s) + \int_{j_{\tilde{\delta}}\tilde{T}_s}^{j_\delta T_s} e^{A(j_\delta T_s - \tau)} B a^{\mathsf{a}}(\tau) d\tau$$

so that the output deviation caused by approximation of $\mathcal{R}$ can be obtained by

$$\begin{aligned} &\|\tilde{y}(j_\delta T_s) - \tilde{y}(j_{\tilde{\delta}}\tilde{T}_s)\|_\infty \\ &\qquad \leq \|C(e^{A(j_\delta T_s - j_{\tilde{\delta}}\tilde{T}_s)} - I) x(j_{\tilde{\delta}}\tilde{T}_s)\|_\infty + \int_{j_{\tilde{\delta}}\tilde{T}_s}^{j_\delta T_s} \|C e^{A(j_\delta T_s - \tau)} B a^{\mathsf{a}}(\tau)\|_\infty d\tau. \end{aligned} \tag{4.2.20}$$

Then, since the input bound $\bar{a}^{\mathsf{a}}$ implies

$$\int_{j_{\tilde{\delta}}\tilde{T}_s}^{j_\delta T_s} \|Ce^{A(j_\delta T_s-\tau)}Ba^{\mathsf{a}}(\tau)\|_\infty d\tau \leq \int_{j_{\tilde{\delta}}\tilde{\mathcal{R}}T_a}^{j_\delta \mathcal{R}T_a} \|Ce^{A(j_\delta T_s-\tau)}B\|_M d\tau p\bar{a}^{\mathsf{a}},$$

the output deviation caused by approximation of the ratio $\mathcal{R}$ is given by

$$\begin{aligned}
&\|\tilde{y}(j_\delta T_s) - \tilde{y}(j_\delta \tilde{T}_s)\|_\infty \\
&\quad \leq \|C(e^{A(j_\delta T_s - j_{\tilde{\delta}}\tilde{T}_s)} - I)x(j_{\tilde{\delta}}\tilde{T}_s)\|_\infty + \int_{j_{\tilde{\delta}}\tilde{\mathcal{R}}T_a}^{j_\delta \mathcal{R}T_a} \|Ce^{A(j_\delta T_s-\tau)}B\|_M d\tau p\bar{a}^{\mathsf{a}} \\
&\quad \leq \|C(e^{A(j_\delta T_s - j_{\tilde{\delta}}\tilde{T}_s)} - I)\|_M \|x(j_{\tilde{\delta}}\tilde{T}_s)\|_1 + \int_{j_{\tilde{\delta}}\tilde{\mathcal{R}}T_a}^{j_\delta \mathcal{R}T_a} \|Ce^{A(j_\delta T_s-\tau)}B\|_M d\tau p\bar{a}^{\mathsf{a}} \\
&\quad \leq \|C(e^{A(j_\delta T_s - j_{\tilde{\delta}}\tilde{T}_s)} - I)\|_M \mathcal{A}_j + \int_{j_{\tilde{\delta}}\tilde{\mathcal{R}}T_a}^{j_\delta \mathcal{R}T_a} \|Ce^{A(j_\delta T_s-\tau)}B\|_M d\tau p\bar{a}^{\mathsf{a}}
\end{aligned}$$

where

$$\mathcal{A}_j := \int_0^{j_\delta T_s} \|e^{A\tau}B\|_1 d\tau p\bar{a}^{\mathsf{a}} \geq \left\| \int_0^{j_{\tilde{\delta}}\tilde{T}_s} e^{A(j_{\tilde{\delta}}\tilde{T}_s-\tau)}Ba^{\mathsf{a}}(\tau)d\tau \right\|_1 = \|x(j_{\tilde{\delta}}\tilde{T}_s)\|_1.$$

This completes the proof. ■

Note that as the truncation level m get larger (approximate accuracy increases), stealthy assurance time t_{sth} becomes longer since $(j_\delta T_s - j_{\tilde{\delta}}\tilde{T}_s) \to 0$ as $m \to \infty$. On the other hand, obviously, lower attack bound $\bar{a}^{\mathsf{a}}$ provides longer stealthy assurance time when the truncation level is fixed. Keeping in mind these facts, by adjusting m and $\bar{a}^{\mathsf{a}}$, the adversary can design the masking attack having practical stealthiness, which secures the stealthy assurance time to the desired extent.

4.3 Simulation Results

4.3.1 Numerical Example: $\mathcal{R} = 1$ with $\delta = 0$

In this subsection, we study a simple example in order to illuminate the attack generation procedure for the case $\mathcal{R} = 1$ without offset, as discussed in Remark

4.2.2. For this, let us consider the error dynamics (4.2.2) with

$$A = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -5 & -3 \\ 0 & 2 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}.$$

With ZOH and sampler whose sampling periods are $T_a = T_s = 1$ sec (and thus $\mathcal{R} = 1$), its sampled-data system is given by (4.2.3) with

$$A_{\mathsf{d}} = \begin{bmatrix} 0.368 & 0 & 0 \\ 0 & -0.121 & -0.257 \\ 0 & 0.171 & 0.306 \end{bmatrix}, \quad B_{\mathsf{d}} = \begin{bmatrix} 0.632 & 0 \\ 0 & 0.086 \\ 0 & 0.231 \end{bmatrix}.$$

Note that the transfer function $C(zI - A_d)^{-1}B_d$ of this sampled-data system has no zeros, hence the zero dynamics attack [TSSJ12, TSSJ15] is not applicable to this system. In contrast, the proposed masking attack can be designed because the conditions of Assumption 4.2.1 hold. Indeed, $q < Rp$ so that (a) holds and the matrix $\mathcal{C}\Pi = CB_{\mathsf{d}}$ has full row rank so that (c) holds. Also, the matrix B_{d} has full column rank, and so, by (the proof of) Proposition 4.2.3, the condition (b) holds with $t_k^* = i^*/\beta = 1/1$.

Under the setting of $H_k = k$, the attack sequence $a^{\mathsf{a}}\langle k \rangle = \kappa_{\langle k \rangle}\eta_{\langle k \rangle} + \zeta_{\langle k \rangle}$ is constructed as follows:

- *Step 1:* Set $\zeta_{\langle 1 \rangle} = \mathrm{col}(0, 0)$ and choose $\eta_{\langle 1 \rangle} = \mathrm{col}(-0.343, 0.939)$ such that

  $$\eta_{\langle 1 \rangle} \in \ker \mathcal{C}\Pi \quad \text{and} \quad \Phi_k^* \eta_{\langle 1 \rangle} = B_{\mathsf{d}}\eta_{\langle 1 \rangle} \neq 0.$$

  Then, select $\kappa_1 = 3.15$ satisfying the inequality $\kappa_1 \|B_{\mathsf{d}}\eta_{\langle 1 \rangle}\| \geq H_1$ (4.2.12) and update $\tilde{x}_1^{\mathsf{c}} = B_{\mathsf{d}}(\kappa_1\eta_{\langle 1 \rangle} + \zeta_{\langle 1 \rangle})$.

- *Step 2:* Choose $\zeta_{\langle 2 \rangle}$ such that (4.2.14) holds (i.e., $CB_{\mathsf{d}}\zeta_{\langle 2 \rangle} = -CA_{\mathsf{d}}\tilde{x}_1^{\mathsf{c}}$). For convenience, let $\eta_{\langle 2 \rangle} = \eta_{\langle 1 \rangle}$ as discussed in Remark 4.2.3. Then, select κ_2 for (4.2.15), and set $\tilde{x}_2^{\mathsf{c}} = A_{\mathsf{d}}\tilde{x}_1^{\mathsf{c}} + B_{\mathsf{d}}(\kappa_2\eta_{\langle 2 \rangle} + \zeta_{\langle 2 \rangle})$.

- *Step 3,$\cdots$:* Similarly, the remaining steps proceed with $\eta_{\langle k \rangle} = \eta_{\langle 1 \rangle}$.

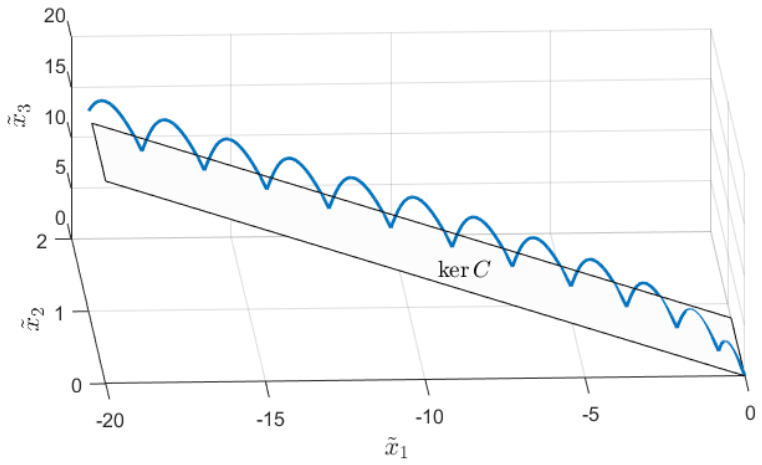


Figure 4.4: Continuous-time state trajectory $\tilde{x}(t)$ (solid blue line) and $\ker C$ (plane).

Fig. 4.4 shows the continuous-time state $\tilde{x}(t)$ from its initial condition $\tilde{x}(0) = \mathrm{col}(0, 0, 0)$ when the constructed attack sequence is injected into the input at $t = 0$ sec. Note that $\tilde{x}(t)$ is the error variable between the state $x(t)$ under attack and the attack-free state $\bar{x}(t)$. In this figure, it is observed that the error $\tilde{x}(t)$ moves far from the origin while it repeatedly encounters $\ker C$, which makes the sampled error output $\tilde{y}(jT_s)$ remain zero as seen in Fig. 4.5 (*zero-stealthy* property). Also, one can see in Fig. 4.6 that the *disruptive property* is effective; that is, $\|\tilde{x}(t_k)\| \geq H_k$ for $(k-1)T_a < t_k \leq kT_a$ (here, $t_k = kT_a$).

To discuss further, we now consider a particular case when the input u_i is saturated so that the maximum magnitude of the attack signal a_i^{a} is necessarily limited. Since the attacker cannot make the input arbitrarily large in this case, the size of the disruption threshold H_k is also naturally limited. Nevertheless, the attacker is still able to construct the masking attack by repeating (off-line) design of attack and its confirmation. Specifically, the attacker first roughly sets the disruption threshold H_k as a bounded sequence (for example, H_k can be taken as constant), and then computes the attack signal via the proposed algorithm. Next, by comparing the magnitude of the obtained attack signal with the saturation level, it is simply verified whether the attack signal is realizable or not. If not, the

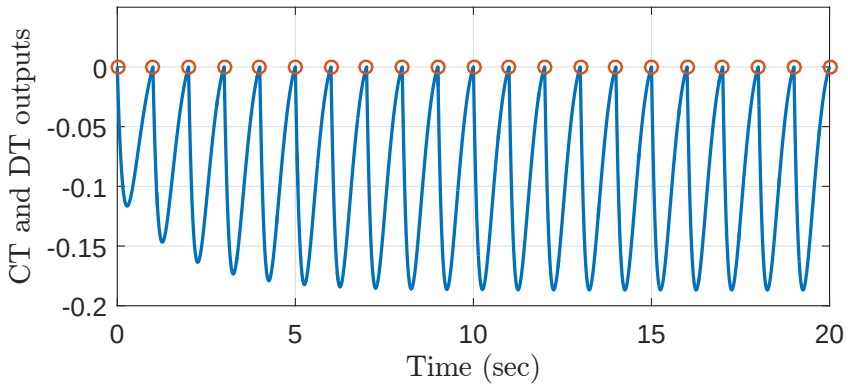


Figure 4.5: Continuous-time output error $\tilde{y}(t)$ (solid blue line) and sampled output error $\tilde{y}(jT_s)$ (red circle).

adversary may repeat the same procedure with smaller H_k, which will bring an attack signal with acceptable magnitude in the end. Fig. 4.7 shows an example where the attack a_i^{a} needs to be saturated as $-30 \leq a_i^{\mathsf{a}} \leq 30$ for all $i = 0, 1, \cdots$. It is seen in the figure that the smaller the constant H_k is, the smaller the maximum magnitude of a_i^{a} is. From this, one can expect that the masking attacks associated with the constant H_k smaller than 5 ensure the additional requirement on the input saturation. We also note that it is rather challenging to carry out explicitly the relation between the maximum magnitude of a_i^{a} and the selection of H_k, which could be one of our further researches.

4.3.2 X-38 Vehicle: $\mathcal{R} = 4$ with $\delta = 0$

As another example, we consider X-38 vehicle model which is a prototype flight test vehicle for crew return [SW00]. The X-38 is operated by a multi-rate digital controller whose holder (or actuator) operates four times faster than the sampler (sensor) with $T_a = 0.04$ sec and $T_s = 0.16$ sec (i.e., $\mathcal{R} = T_s/T_a = 4$). The continuous-time system has 3 inputs, 9 outputs, and 11 states ($A \in \mathbb{R}^{11\times 11}$, $B \in \mathbb{R}^{11\times 3}$, and $C \in \mathbb{R}^{9\times 11}$). More detailed information on the X-38 is provided in [SW00], [BS98].

From the information of the X-38 model in [SW00] (that is omitted here), one

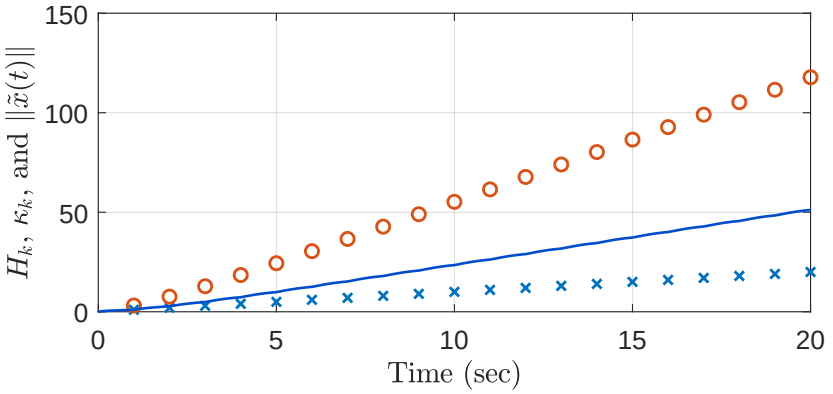


Figure 4.6: Sequence H_k (blue cross), selected κ_k (red circle), and $\|\tilde{x}(t)\|$(blue solid line).

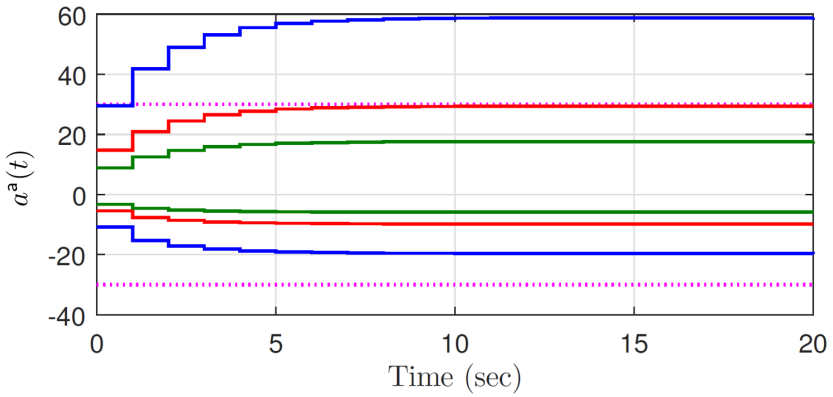


Figure 4.7: Attack signal $a^{\mathsf{a}}(t) \in \mathbb{R}^2$ with $H_k = 10$ (blue line), $H_k = 5$ (red line), $H_k = 3$ (green line), and input saturation ± 30 (dotted line).

can verify that Assumption 4.2.1 holds by the following reasons:

- $Rp = 12$ and $q = 9$, and so, the condition (a) holds (i.e., $\mathcal{C}\Pi \in \mathbb{R}^{9\times 12}$ so that $\ker \mathcal{C}\Pi \neq \{0\}$),

- the matrix B_{d} has full column rank, and there exists a non-zero vector z such that $\mathcal{C}\Pi z = 0$ where the first 3 components are a non-zero vector in $\mathbb{R}^3$. Then, by the proof of Proposition 4.2.3, $i^* = 1$. Therefore, the condition

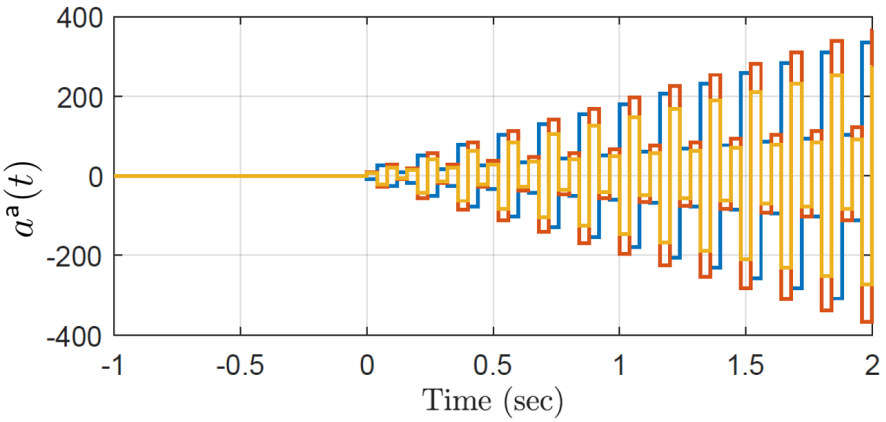


Figure 4.8: Generated attack signal $a^{\mathsf{a}}(t)$.

(b) holds with $t_k^* = i^*/\beta = 1/4$,

- the matrix $\mathcal{C}\Pi$ has full row rank so that im $\mathcal{C}\Pi = \mathbb{R}^9$ and the condition (c) holds.

Now, following the proposed attack generation procedure, we construct an attack sequence $a^{\mathsf{a}}\langle k\rangle = \kappa_k \zeta_{\langle k\rangle} + \eta_{\langle k\rangle}$ with disruptive property $H_k = 0.5k$. In particular, we choose $\eta_{\langle k\rangle}$ as follows:

$$\eta_{\langle k\rangle} = \operatorname{col}\left(\begin{bmatrix}-0.132\\0.145\\0.108\end{bmatrix}, \begin{bmatrix}0.397\\-0.434\\-0.324\end{bmatrix}, \begin{bmatrix}-0.396\\0.434\\0.324\end{bmatrix}, \begin{bmatrix}0.132\\-0.144\\-0.108\end{bmatrix}\right)$$
$$\in \ker \mathcal{C}\Pi \text{ and } \Phi_k^* \eta_{\langle k\rangle} \neq 0,$$

and $\zeta_{\langle k\rangle}$, κ_k are selected to satisfy (4.2.16) and (4.2.17), respectively.

The constructed attack sequence $a^{\mathsf{a}}(t)$ is demonstrated in Fig. 4.8, and Fig. 4.9 illustrates the state error $\tilde{x}(t)$ when the attack is injected into the system at $t = 0$ sec. While the measured output at each sampling time looks normal (Fig. 4.10.(c)), the error states $\tilde{x}(t)$ are fluctuating as seen in Fig. 4.9, and likewise, the continuous-time output $\tilde{y}(t)$ is also not calm as seen in Fig. 4.10.(b) (the attack-free continuous-time output $\bar{y}(t)$ is depicted in Fig. 4.10.(a) for comparison).

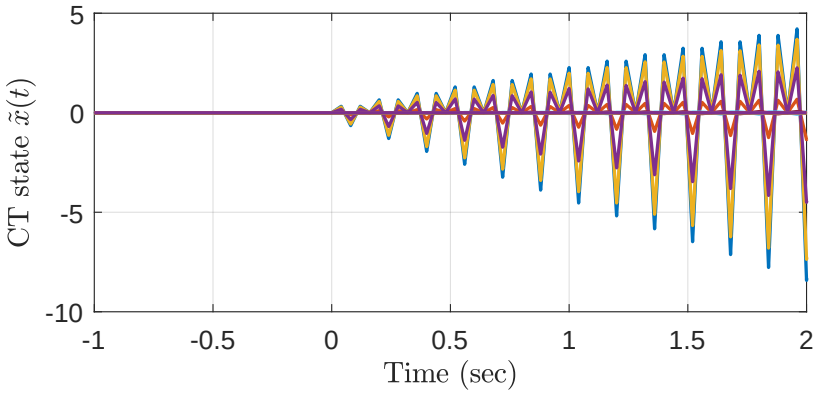


Figure 4.9: Continuous-time error state $\tilde{x}(t)$ of the X-38 model.

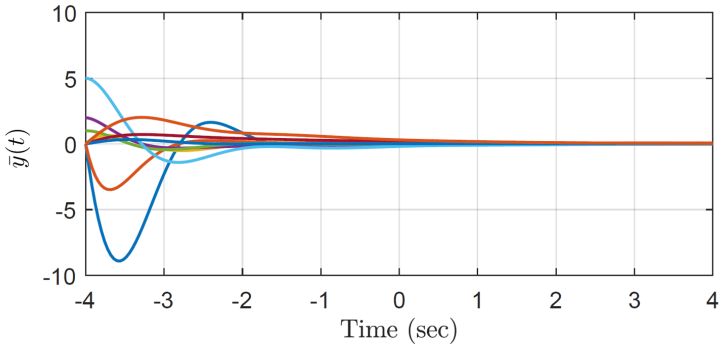
4.3.3 Numerical Example: $\mathcal{R} = 0.4$ with $\delta = 0.75$

In this subsection, we show that the proposed attack is effective under Assumption 4.2.1, even if the sampling period of the sensor is shorter than that of the actuator and there exists an offset δ (this cannot be dealt with in [KPSE16]). Specifically, let us consider the case where $T_a = 1$ sec and $T_s = 0.4$ sec, so that $\mathcal{R} = 0.4/1 = 2/5 = \beta/\alpha$ (i.e., there are 5 sensings and 2 actuations for each cluster). Moreover, let us assume an offset, $\delta = 0.3/0.4 = 0.75$ (i.e., the sensor starts 0.3 sec later than the actuator). The considered plant is described by a minimal realization of

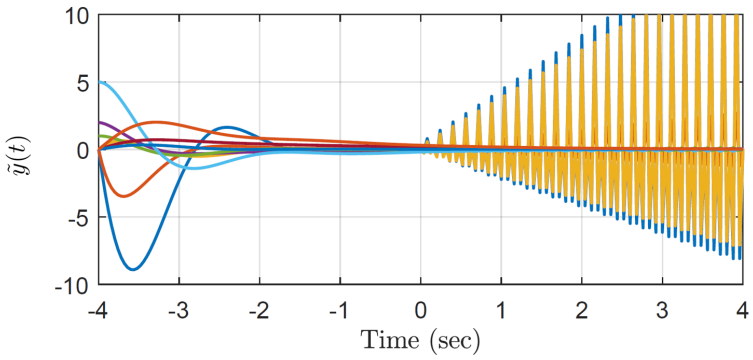
$$G(s) = \begin{bmatrix} \dfrac{1}{s+1} & \dfrac{2}{(s+2)(s+3)} & \dfrac{4}{(s+4)(s+5)} \end{bmatrix}.$$

From the minimal realization $A \in \mathbb{R}^{5\times5}$, $B \in \mathbb{R}^{5\times3}$, and $C \in \mathbb{R}^{1\times5}$, one can verify Assumption 4.2.1 as follows:

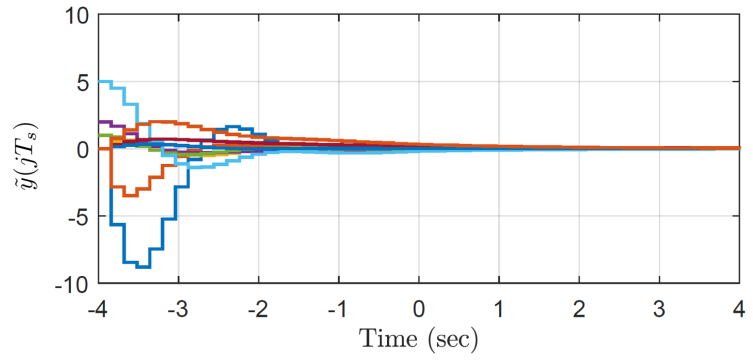
- The plant has 3 inputs, 1 output, and $\mathcal{R} = 0.4$. Hence, $q = 1 < Rp = 1.2$, and so, the condition (a) holds (i.e., $\mathcal{C}\Pi \in \mathbb{R}^{5\times6}$ so that $\ker \mathcal{C}\Pi \neq \{0\}$).
- The matrix B_{d} has full column rank, and there exists a non-zero vector z such that $\mathcal{C}\Pi z = 0$ where the first 3 components are a non-zero vector in $\mathbb{R}^3$. Then, by the proof of Proposition 4.2.3, $i^* = 1$. Therefore, the condition (b) holds with $t_k^* = i^*/\beta = 1/2$.



(a) Continuous-time output $\bar{y}(t)$ under no attack.



(b) Continuous-time output $\tilde{y}(t)$ under the proposed attack.



(c) Discrete-time output $\tilde{y}(jT_s)$ under the proposed attack.

Figure 4.10: Continuous-time and discrete-time outputs with and without attack.

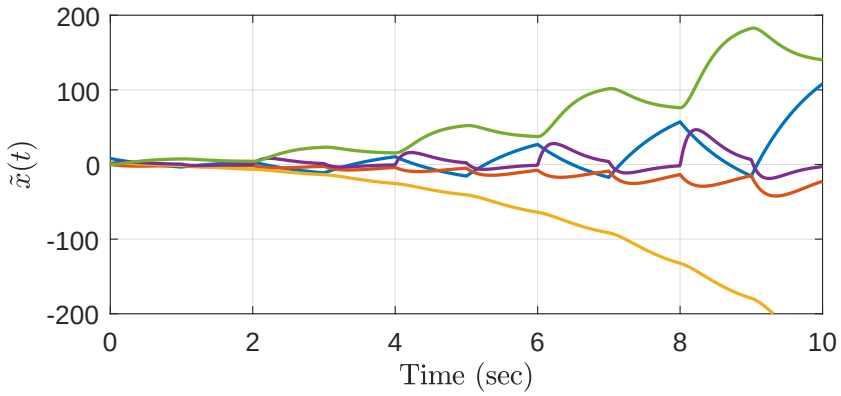


Figure 4.11: Behavior $\tilde{x}(t)$ of error dynamics.

- The matrix $\mathcal{C}\Pi$ has full row rank so that im $\mathcal{C}\Pi = \mathbb{R}^5$ and the condition (c) holds.

An attack sequence a_i^{a} with $H_k = 10k$ is constructed by following the proposed procedure. Specifically, $\eta_{\langle k \rangle}$ have chosen as

$$\eta_{\langle k \rangle} = \mathrm{col}(-0.188, -0.163, 0.746, 0.138, -0.467, 0.379)$$

for all $k \geq 1$, which satisfies $\eta_{\langle k \rangle} \in \ker \mathcal{C}\Pi$ and $\Phi_k^* \eta_{\langle k \rangle} \neq 0$. The quantities κ_k and $\zeta_{\langle k \rangle}$ are selected appropriately by the attack generation procedure in Section 4.2.

The simulation results illustrate the behavior of $\tilde{x}(t)$ in Fig. 4.11, and the output signal $\tilde{y}(t)$ in Fig. 4.12, respectively. It is seen that, even if the error variable $\tilde{x}(t)$ and the continuous-time output $\tilde{y}(t)$ diverge, the output measurements (represented as red circles in Fig. 4.12) remain zero, so that both stealthiness and disruptive property are achieved.

Remark 4.3.1. In practice, the adversary may encounter various sources of uncertainties: for instance, an input delay exists in the network communication (so that Δ is not exactly known); or the ratio $\mathcal{R} = T_s/T_a$ is perturbed. Even in these cases, as long as the uncertain quantity is small and the time interval of attacker's interest is finite, the masking attack under imperfect knowledge could remain stealthy in a practical sense, from which the detection of the attack is de-

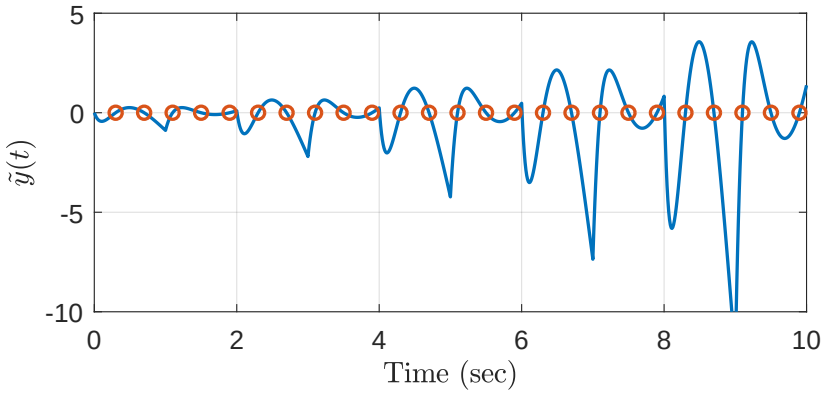


Figure 4.12: Output error $\tilde{y}(t)$ (blue solid line) and its sampled measurements $\tilde{y}(j_\delta T_s)$ with offset $\delta = 0.75$ (red circle).

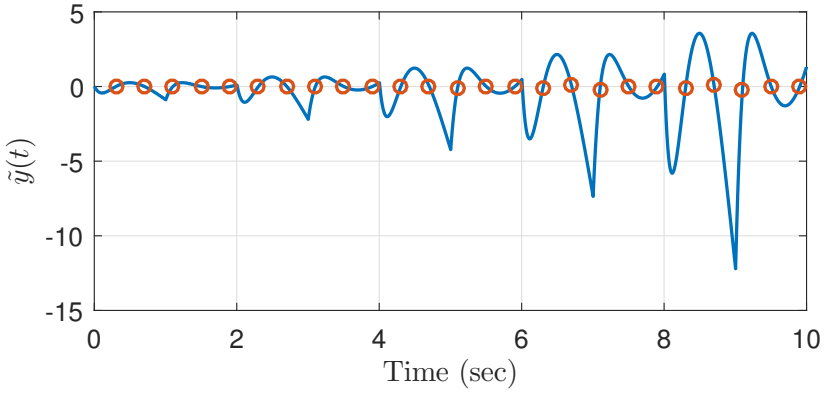


Figure 4.13: Output error $\tilde{y}(t)$ (blue solid line) and its sampled measurements $\tilde{y}(j_\delta T_s)$ (red circle) when $T_s = 0.4$ sec and $T_a = 1$ sec with input delay 0.004 sec. The injected attack is designed assuming that $T_s = 0.4$ sec and $T_a = 1$ sec.

layed enough until a fatal damage is incurred in the plant. To verify this, we provide two simulations in Figs. 4.13 and 4.14. In both simulations, the attack signal is supposed to be designed using the nominal parameters $\Delta = 0.3$ sec, $T_s = 0.4$ sec, and $T_a = 1$ sec. Fig. 4.13 depicts the case when the actual input signal is delayed by 0.004 sec (and thus the actual Δ is given by $\Delta = 0.296$ sec). On the other hand, in Fig. 4.14, it is assumed that the actual sensing period is slightly

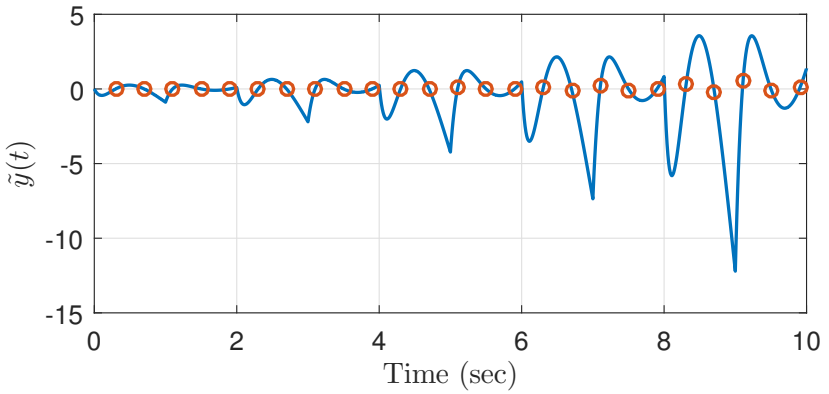


Figure 4.14: Output error $\tilde{y}(t)$ (blue solid line) and its sampled measurements $\tilde{y}(j_\delta T_s)$ (red circle) when $T_s = 0.4004$ sec, $T_a = 1$ sec, while the attack is designed assuming that $T_s = 0.4$ sec and $T_a = 1$ sec.

distorted as $T_s = 0.4004$ sec. In both scenarios, one can see that the sampled output remains near zero (i.e., the attack is practically stealthy in a sense), while the continuous-time output deviates largely from the normal behavior. □

Remark 4.3.2. In terms of a defender against the masking attack, it is tricky to detect or incapacitate the attack. This is because the masking attack is constructed based on the assumption of perfect knowledge of the system dynamics and uses geometric weak points (i.e., Assumption 4.2.1) to secure stealthiness and lethality of the attack. Thus, to prevent this attack, the defender should modify the system confidentially or Assumption 4.2.1 should be broken; it is literally a duel of the shield and spear. We introduce two simple strategies for enhancing shield. As a confidential modification, we can modify the holding device (e.g., the generalized hold), by which the input matrix B_{d} is changed so that the masking attack cannot be stealthy without the information of the holding device. On the other hand, if the defender can obtain inter-sample information of measurements so that Assumption 4.2.1 –(a) is violated with additional measurements, the attacker cannot construct the masking attack having zero-stealthy property. □

Chapter 5

Conclusion of Dissertation

The main objective of this dissertation is to address new vulnerabilities and countermeasures against them on cyber-physical systems, which are originated from the properties of sampled-data systems. The details are listed as follows.

i) In Chapter 2, we have investigated the zero dynamics attack that utilizes unstable sampling zeros, which could often appear even for minimum phase continuous-time systems when a digital control is used. Noted by the effectiveness of the zero dynamics attack being originated from the instability of discrete-time zeros, we proposed to replace the zero-order hold at the actuator block by the generalized hold as a new countermeasure against the zero dynamics attack. By employing the generalized hold and design appropriate hold gains, the discrete-time zeros can be shifted into the stable region so that the effect of the attack becomes negligible. Then, we examined the trade-off that comes from employing the generalized hold function. That is, the degradation (or fluctuation) in the inter-sample behavior occurs since in some cases, the generalized hold requires high gain in the subintervals to change the discrete-time transfer function. In order to reflect this trade-off into the design, we formulated an optimization problem for which the performance index is closely related to the inter-sample behavior, and showed that the problem can be converted into a convex problem. Furthermore, we carried out an experiment with a DC/DC converter combined with a high-order low-pass filter (electrical circuit) to verify the proposed theoretical results.

ii) In Chapter 3, we studied the zero-dynamics attack that invades through the sensor network. We showed that it is undetectable from the anomaly detector and effective as long as the continuous time plant is unstable. Subsequently, as a countermeasure against this attack, we proposed to install the generalized hold feedback which connects the discrete-time output and the continuous-time input, by which the poles of the sampled-data system can be stabilized. By doing this, the zero-dynamics sensor attack can be prevented fundamentally since no matter what the attacker's knowledge is, the attack either is detected or becomes negligible. Although the zero-dynamics sensor attack is successfully prevented by the proposed strategy, employing the generalized hold feedback may cause an unfavorable inter-sample behavior (because of fluctuating hold gain). To attenuate this drawback, we employed an optimal controller where a discrete-time linear quadratic regulator minimizing a continuous-time performance index for reflecting the inter-sample behavior.

iii) In Chapter 4, apart from the threat targeting the unstable zeros of discrete-time systems, we discovered that another type of stealthy but disruptive attack is also possible, if there exists enough input resources compared to the output in the multi-rate and multi-input sense. Specifically, when the sampling ratio between the zero-order hold and sensor is allowed to be rational numbers for MIMO systems, we established the concept of clustering the time frame and then computed a lifted system corresponding to such clusters. Then, by taking a closer look at the state trajectory in both continuous- and discrete-time domains, we showed that how the additional input resources and full system knowledge enable the adversary to compromise the inter-sample behavior of the sampled-data system, while being perfectly undetected at each sampling time.

BIBLIOGRAPHY

[ACS09] S. Amin, A. A. Cardenas, and S. S. Sastry. *Hybrid Systems: Computation and Control: Safe and secure networked control systems under denial-of-service attacks.* Springer-Verlag, 2009.

[AENR88] M. M. Akhter A. E. Naeini and S. M. Rock. Effect of model uncertainty on failure detection: the threshold selector. *IEEE Transactions on Automatic Control*, 33, 1988.

[AMAY+18] M. N. Al-Mhiqani, R. Ahmad, W. Yassin, A. Hassan, Z. Z. Abidin, N. S. Ali, and K. H. Abdulkareem. Cyber-security incidents: a review cases in cyber-physical systems. *International Journal of Advanced Computer Science and Applications*, 9, 2018.

[BKL+17] J. Back, J. Kim, C. Lee, G. Park, and H. Shim. Enhancement of security against zero dynamics attack via generalized hold. In *Proceedings of 56th Annual IEEE Conference on Decision and Control*, pages 1350–1355, 2017.

[BS98] J. Bain and J. Sunkel. Autonomous control for subsonic flight of the x-38. In *Proceedings of Guidance, Navigation, and Control Conference and Exhibit*, pages 909–922, 1998.

[Con10] J. P. Conti. The day the samba stopped. *Engineering and Technology*, 5:46–47, 2010.

[DB08] R. C. Dorf and R. H. Bishop. *Modern Control Systems* Twelfth Ed. Pearson, 2008.

[FD94] P. M. Frank and X. Ding. Frequency domain approach to optimally robust residual generation and evaluation for model-based fault diagnosis. *Automatica*, 30, 1994.

[FG94] A. Feuer and G. C. Goodwin. Generalized sample hold functions–frequency domain analysis of robustness, sensitivity, and intersample difficulties. *IEEE Transactions on Automatic Control*, 39, 1994.

[FG96] A. Feuer and G. C. Goodwin. *Sampling in Digital Signal Processing and Control*. MA: Birkhäuser, Boston, 1996.

[FH02] H. Fujimoto and Y. Hori. High-performance servo systems based on multirate sampling control. *Control Engineering Practice*, 10:773–781, 2002.

[FKK03] H. Fujimoto, F. Kawakami, and S. Kondo. Multirate repetitive control and applications. In *Proceedings of American Control Conference*, pages 2875–2880, 2003.

[FQCZ20] C. Fang, Y. Qi, P. Cheng, and W. X. Zheng. Optimal periodic watermarking schedule for replay attack detection in cyber–physical systems. *Automatica*, 112, 2020.

[FTD14] H. Fawzi, P. Tabuada, and S. Diggavi. Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE Transaction Automatic Control*, 59:1454–1467, 2014.

[HA02] T. Hagiwara and M. Araki. Design of a stable state feedback controller based on the multirate sampling of the plant output. *IEEE Transactions on Automatic Control*, 33:812–819, 2002.

[HFA90] T. Hagiwara, T. Fujimura, and M. Araki. Generalized multirate-output controllers. *International Journal of Control*, 42:597–612, 1990.

[HS19] J. Ha and H. Shim. Study on realizable generalized hold functions as a countermeasure against zero dynamics attack. In *Proceedings of*

58th IEEE Conference on Decision and Control, pages 5362–5367, 2019.

[HZ16] A. Hoehn and P. Zhang. Detection of covert attacks and zero dynamics attacks in cyber-physical systems. In *Proceedings of American Control Conference*, pages 302–307, 2016.

[Kab87] P. Kabamba. Control of linear systems using generalized sampled-data hold functions. *IEEE Transactions on Automatic Control*, 32:772–783, 1987.

[Kar11] S. Karnouskos. Stuxnet worm impact on industrial cyber-physical system security. In *Proceedings of 37th Annual Conference of the IEEE Industrial Electronics Society*, pages 4490–4494, 2011.

[KBP+20] J. Kim, J. Back, G. Park, C. Lee, H. Shim, and P. G. Voulgaris. Neutralizing zero dynamics attack on sampled-data systems via generalized holds. *Automatica*, 113, 2020.

[Kha02] H. K. Khalil. *Nonlinear Systems,* Third Ed. Prentice-Hall, Upper Saddle River, NJ, 2002.

[KKS17] A. Khazraei, H. Kebriaei, and F. R. Salmasi. A new watermarking approach for replay attack detection in lqg systems. In *Proceedings of 56th Annual Conference on Decision and Control (CDC)*, pages 5143–5148, 2017.

[KPSE16] J. Kim, G. Park, H. Shim, and Y. Eun. Zero-stealthy attack for sampled-data control systems: the case of faster actuation than sensing. In *Proceedings of 55th Annual Conference on Decision and Control*, pages 5956–5961, 2016.

[KS72] H. Kwakernaak and R. Sivan. *Linear optimal control systems.* New York, Wiley-interscience, 1972.

[KU08] K.Lavanya and B. Umamaheswari. Design of digital multi-rate controller using frequency domain analysis. *Journal of Circuits, Systems, and Computers*, 17:675–684, 2008.

[LAC16] R. M. Lee, M. J. Assante, and T. Conway. Analysis of the cyber attack on the Ukrainian power grid, 2016.

[Lan11] R. Langner. Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security Privacy Magazine*, 9:49–51, 2011.

[LNR11] Y. Liu, P. Ning, and M. K. Reiter. False data injection attacks against state estimation in electric power grids. *ACM Transactions on Information and System Security*, 14:1–33, 2011.

[lSAQ07] M. De la Sen and S. Alonso-Quesada. Model matching via multirate sampling with fast sampled input guaranteeing the stability of the plant zeros: extensions to adaptive control. *IET Control Theory and Applications*, 1:210–225, 2007.

[LSE15] C. Lee, H. Shim, and Y. Eun. A secure and robust state estimation under sensor attacks, measurement noise, and process disturbances: observer-based combinatorial approach. In *Proceedings of the 14th European Control Conference*, pages 1866–1871, 2015.

[LSE19] C. Lee, H. Shim, and Y. Eun. On redundant observability: From security index to attack detection and resilient state estimation. *IEEE Transactions on Automatic Control*, 64:775–782, 2019.

[MA18] Y. Mao and E. Akyol. Detectability of cooperative zero-dynamics attack. In *Proceedings of 56th Annual Allerton Conference on Communication, Control, and Computing*, pages 227–234, 2018.

[MB12] J. Mattingley and S. Boyd. CVXGEN: a code generator for embedded convex optimization. *Optimization and Engineering*, 13:1–27, 2012.

[Mor84] T. Mori. Note on the absolute value of the roots of a polynomial. *IEEE Transactions on Automatic Control*, 29:54–55, 1984.

[MS09] Y. Mo and B. Sinopoli. Secure control against replay attacks. In *Proceedings of 47th Annual Allerton Conference on Communication, Control, and Computing*, pages 911–918, 2009.

[MS10] Y. Mo and B. Sinopoli. False data injection attacks in control systems. In *Proceedings of Workshop on Secure Control Systems*, 2010.

[MTO07] M. Mizuochi, T. Tsuji, and K. Ohnishi. Multirate sampling method for acceleration control system. *IEEE Transactions on Industrial Electronics*, 53:1462–1471, 2007.

[NHV15] M. Naghnaeian, N. Hirzallah, and P. G. Voulgaris. Dual rate control for security in cyber-physical systems. In *Proceedings of 54th IEEE Conference on Decision and Control*, pages 1415–1420, 2015.

[NHV19] M. Naghnaeian, N. Hirzallah, and P. G. Voulgaris. Security via multirate control in cyber–physical systems. *System & Control Letters*, 124:12–18, 2019.

[PLS18] G. Park, C. Lee, and H. Shim. On stealthiness of zero-dynamics attacks against uncertain nonlinear systems: a case study with quadruple–tank process. In *Proceedings of International Symposium on Mathematical Theory of Networks and Systems*, pages 10–17, 2018.

[PLS+19] G. Park, C. Lee, H. Shim, Y. Eun, and K. H. Johansson. Stealthy adversaries against uncertain cyber-physical systems: Threat of robust zero-dynamics attack. *IEEE Transactions on Automatic Control*, 64:4907–4919, 2019.

[PSL+16] G. Park, H. Shim, C. Lee, Y. Eun, and K. H. Johansson. When adversary encounters uncertain cyber-physical systems: robust zero-dynamics attack with disclosure resources. In *Proceedings of 55th IEEE Conference on Decision and Control*, pages 5085–5090, 2016.

[sHS84] K. J. Åström, P. Hagander, and J. Sternby. Zeros of sampled systems. *Automatica*, 20:31–38, 1984.

[SM07] J. Slay and M. Miller. Lessons learned from the maroochy water breach. *Critical Infrastructure Protection*, 253:73–82, 2007.

[SPY+17] H. Sun, C. Peng, T. Yang, H. Zhang, and W. He. Resilient control of networked control systems with stochastic denial of service attacks. *Neurocomupting*, 270:170–177, 2017.

[SSP03] C. Fantuzzi S. Simani and R. J. Patton. *Model-based fault diagnosis techniques*. Springer, 2003.

[Sta11] CNN Wire Staff. Obama says u.s. has asked iran to return drone aircraft, 2011.

[SW00] L. S. Shieh and W. M. Wang. Design of lifted dual-rate digital controllers for x-38 vehicle. *Journal of Guidance, Control, and Dynamics*, 23:629–639, 2000.

[TSSJ12] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson. Revealing stealthy attacks in control systems. In *Proceedings of 50th Annual Allerton Conference on Communication, Control, and Computing*, pages 1806–1813, 2012.

[TSSJ15] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson. A secure control framework for resource-limited adversaries. *Automatica*, 51:135–148, 2015.

[WLS17] S. Weerakkody, X. Liu, and B. Sinopoli. Robust structural analysis and design of distributed control systems to prevent zero dynamics attacks. In *Proceedings of 56th Annual Conference on Decision and Control*, pages 1356–1361, 2017.

[WS15] S. Weerakkody and B. Sinopoli. Detecting integrity attacks on control systems using a moving target approach. In *Proceedings of 54th Annual Conference on Decision and Control*, pages 5820–5826, 2015.

[YG14] J. I. Yuz and G. C. Goodwin. *Sampled-Data Models for Linear and Nonlinear Systems*. Springer-Verlag, 2014.

[YT01] Z. J. Yang and M. Tateishi. Adaptive robust nonlinear control of a magnetic levitation system. *Automatica*, 37:1125–1131, 2001.

[ZCSC16] H. Zhang, P. Cheng, L. Shi, and J. Chen. Optimal dos attack scheduling in wireless networked control system. *IEEE Transactions on Control Systems Technology*, 24:843 – 852, 2016.

국문초록

Design and Countermeasure of Stealthy Attacks on Cyber-Physical Systems in Sampled-data Framework

샘플 데이터로 표현되는 사이버-물리 시스템의 취약점 분석 및 검출 불가능한 공격에 대한 방어 기법

디지털 장치들의 연산 속도와 네트워크 전송 속도의 급진적인 발전으로 고전적인 제어 시스템이 네트워크를 통해 원격으로 제어되는 사이버-물리 시스템(cyber-physical systems)이 등장하기 시작했다. 이러한 사이버-물리 시스템은 제어기와 제어 대상의 분리라는 특성상 외부의 악의적인 공격신호로 부터 공격당할 수 있는 잠재적인 위험에 노출되어 있으며 파워플랜트의 원격감시제어(SCADA, Supervisory Control And Data Acquisition)와 같은 사회 기반 시설과도 밀접한 연관이 있어 그 보안성에 관한 연구의 필요성이 강조되고 있다. 본 논문은 사이버-물리 시스템이 연속시간으로 이루어진 물리 플랜트(physical plant)와 디지털 제어기로 이루어져 있다는 사실로부터 이를 영차홀드(zero-order hold)와 샘플러(sampler)로 이산화(discretize)되는 샘플-데이터 시스템으로 표현하고, 연속시간과 이산시간의 결합으로 부터 발생할 수 있는 사이버 공격에 대한 이론적인 취약점을 분석하고 그에 대한 해결책을 제시한다.

구체적으로 본 논문에서는 다음의 세 가지 주제들을 다룬다. 첫 번째로, 본 논문은 시스템의 불안정한(unstable) 영점(zero)의 정보를 이용하여 입력 네트워크를 통해 주입될 경우 검출불가능(undetectable)한 영동역학 공격(zero dynamics attack)이 샘플 데이터 시스템에서 발생하는 샘플링 영점(sampling zero)을 이용하여도 가능하다는 점을 밝힌다. 그리고 영차홀드 대신 일반화된 홀드(generalized hold)를 이용할 경우 이산시간 시스템의 이산시간 영점을 모두 안정한(stable)한 영역으로 할당할 수 있다는 사실에 근거하여 영동역학 공격에 대한 근본적인 대응책으로 영차홀드를 일반화된 홀드로 대체하는 방안을 제안한다. 추가적으로, 일반화된 홀드를 이용할 경우 발생하는 성능저하를 최소화 하기 위해 볼록(convex) 최적화 문제로 일반화된 홀드를 설계하는 방법을 제시한다. 다른 한편, 이산시간 시스템의 출력 센서 네트워크를 입력 그리고 고장 검출기(fault detector)의 잔여신호(residual)를 출력으로 하는 시스템의 영동역학을 이용하여 검출 불가능한 센서 공격이 가능함을 보이고, 이에 대한 해결책으로 이산시간 출력 부터 연속시간 입력까지 일반화된 홀드를 이용한 피드백 루프를 추가하여

공격의 효과를 무효화하는 방법을 제안한다. 또한 이러한 피드백 루프로 인한 제어 성능 저하를 최소화하기 위해 연속시간 비용함수를 최소화하는 이산시간 최적 제어기법의 이용을 제안한다. 마지막으로, 영차홀드와 샘플러의 동작주기가 같지 않은 다중입출력(MIMO) 샘플-데이터 시스템을 쌓인 시스템(lifted system)으로 표현쌓을 때 출력대비 입력 여유분이 많을 경우, 입력 네트워크를 통하여 검출 불가능한 공격을 가능하게 하는 충분조건을 찾고, 이를 활용하여 공격신호를 생성하는 설계법을 제안한다.

주요어: 사이버-물리 시스템, 네트워크 제어시스템, 샘플-데이터 시스템, 영동역학 공격, 일반화된 홀드
학 번: 2014–21659