Security in Cyber-Physical Systems: Controller Design Against Known-Plaintext Attack

Ye Yuan* and Yilin Mo†

Abstract— **A substantial amount of research on the security of cyber-physical systems assumes that the physical system model is available to the adversary. In this paper, we argue that such an assumption can be relaxed, since even if the physical system model is unknown, the adversary might still be able to identify it by observing the control input and sensory data from the system. In such a setup, the attack with the goal of identifying the system model using the knowledge of input-output data can be categorized as a Known-Plaintext Attack (KPA) in the information security literature. We first prove a necessary condition and a sufficient condition, under which the adversary can successfully identify the transfer function of the physical system. Furthermore, we design an algorithm, which is based on spectral factorization, for the adversary to numerically compute the physical system model. We then provide a low-rank controller design which renders the system unidentifiable to the adversary, while trading off the LQG performance. Finally, a numerical example is provided to illustrate the effectiveness of the proposed controller design.**

I. Introduction

Cyber-Physical Systems (CPSs) refer to the embedding of widespread sensing, networking, computation, and control into physical spaces with the goal of making them safer, more efficient and reliable. Driven by the miniaturization and integration of sensing, communication, and computation in cost effective devices, CPSs are bound to transform several industries such as aerospace, transportation, built environment, energy, health-care, and manufacturing, to name a few. While the use of dedicated communication networks has so far sheltered systems from the outside world, use of off-the-shelf networking and computing, combined with unattended operation of a plethora of devices, provides several opportunities for malicious entities to inject attacks on CPSs. A wide variety of motivations exist for launching an attack on CPSs, ranging from economic reasons such as drawing a financial gain, all the way to terrorism. Any attack on safety-critical CPSs may significantly hamper the economy and lead to the loss of human lives. While the threat of attacks on CPSs tends to be underplayed at times, the Stuxnet worm provided a clear sample of the future to come [1], [2].

A substantial amount of research effort has been dedicated to identifying possible security vulnerabilities of the CPS and develop countermeasures. To this end, many attack models, such as stealthy attack[1] [3], [4], [5], [6], [7], replay attack [8], [9] and covert attack [10], have been proposed by various researchers. Teixeira et al. [11] propose a characterization of different attack models based on the attacker's resources, which are divided into three different categories: knowledge of the system model, knowledge of the real-time control and sensory data (disclosure resources) and the capability to modify the control and sensory data (disruptive resources). Their results illustrate that many attack models proposed in the literature require the knowledge of the system models from the adversary. For example, in the stealthy attack scenario [5], the adversary will inject an external control input to the physical system and then remove the physical system's response to this malicious input from the sensors' measurements. The system operator will not be able to detect the attack if the response to the malicious control input is removed perfectly. However, such an attack requires the adversary to know the perfect model of the physical system, which may be difficult to acquire in many practical scenarios, since the modelling information is usually stored inside the controller. On the other hand, we argue that in many situations, the control and sensory data are much easier to acquire. This is due to the fact that these data are typically not encrypted for many CPSs [12]. Furthermore, even if the control and sensory data are encrypted, it might be easier to break the security of sensors and actuators due to their low computational capability. Thus, for the adversary, the disclosure resources may be more available than the model knowledge.

In this paper, we discuss whether the adversary can use its disclosure resources to gain the model knowledge by the means of system identification. We model the CPS as a linear feedback control system, which is illustrated in Fig 1. The adversary is assumed to *only use* its disclosure resources. In other words, it can only passively observe the control input u and the sensory data y and cannot inject any disturbances to the system. The goal of the adversary is to learn the physical system model $\mathcal{G}(z)$, which further enables the adversary to launch other attacks, such as stealthy attack and covert attack.

Such an attack model is very similar to the Known-Plaintext Attack (KPA) studied in information security, where the adversary has samples of both the plaintext and the corresponding ciphertext and want to deduce the encryption key. For our case, one can view the system model, the control input u and the sensory data y as the encryption key, plaintext and ciphertext respectively.

It is worth mentioning that with additional disruptive resources, the adversary can also launch a more powerful Chosen-Plaintext Attack (CPA), where it can actively modify

*: Control Group, Department of Engineering, University of Cambridge (Darwin College), UK. Email: yy311@cam.ac.uk.

†: Control & Dynamical Systems, California Institute of Technology, USA. Emails: yilinmo@caltech.edu.

[1]The stealthy attack is also referred to as false data injection attack, zero dynamics attack in the literature.

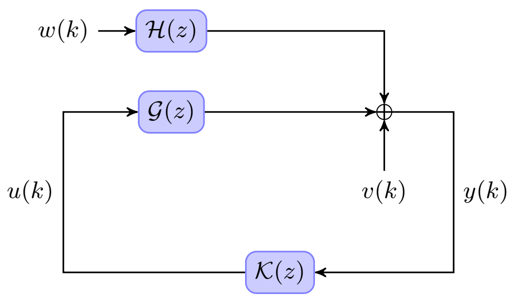

Fig. 1. A general diagram of the CPS. In particular, we consider a widely-used LQG framework in this paper. $\mathcal{G}(z)$ represents the plant while $\mathcal{K}(z)$ the controller.

the control input u and observe the corresponding system output y. However, if the attacker changes the control input u carelessly, it may result in a substantial change in the sensor measurement y, which could enable the system to detect the presence of the malicious third party. If the stealthiness of the attack is of concern to the attacker, then a reasonable strategy of the adversary is to first launch a passive KPA without risking being detected. After a coarse system model is learned, the attacker can then design a stealthy control input u to identify a more accurate model. In such a scenario, the KPA is the first step for the adversary to gain model knowledge.

As a result, we will focus on KPA in this paper. The main contributions of the paper are twofold:

1) We provide a necessary condition and a sufficient condition, under which the system is vulnerable to KPA, i.e., the adversary can successfully identify the system model $\mathcal{G}(z)$. The results can be viewed as an application of classical system identification [13], [14], [15], [16], [17], [18] for the closed-loop system described in Section III.
2) We design a countermeasure to KPA by using a "low-rank" controller design strategy for $\mathcal{K}(z)$ while trading off the LQG control performance.

The rest of the paper is organized as follows: In Section II, we model the system as a linear feedback control system subject to Gaussian process and measurement noise. In Section III, we provide necessary and sufficient conditions, under which the adversary can identify the system model $\mathcal{G}(z)$. We further provide a numerical algorithm for the adversary to compute $\mathcal{G}(z)$. In Section IV, we present a controller design which is resilient to KPA while only incurring minimal control performance loss. The efficiency of the proposed control design is illustrated via a numerical example in Section V. Finally, Section VI concludes the paper.

Notations

$A \succeq B$: $A - B$ is a positive semi-definite matrix. $\mathbb{E}$: expected value. $\mathbb{S}^n$: the set of $n \times n$ symmetric matrices. If U is a positive semidefinite matrix, then $U^{1/2}$ is a positive semidefinite matrix that satisfies $U^{1/2}U^{1/2} = U$. We will use calligraphic letters to denote transfer matrices and normal letters to denote constant matrices. A rational transfer function is called to be *proper* if the degree of the numerator does not exceed the degree of the denominator. It is called strictly proper if the degree of the numerator is less than the degree of the denominator. For a rational transfer matrix $\mathcal{V}(z)$, we define $\mathcal{V}^*(z) = \mathcal{V}^T(\frac{1}{z})$.

II. System model

We model the physical system has a linear time invariant system, which takes the following form:

$$x(k+1) = Ax(k) + Bu(k) + w(k), \tag{1}$$

$$y(k) = Cx(k) + v(k), \tag{2}$$

where $x(k) \in \mathbb{R}^n$, $u(k) \in \mathbb{R}^p$, $y(k) \in \mathbb{R}^m$ are the state, the control input and the sensor measurement at time k respectively. $w(k) \in \mathbb{R}^n$, $v(k) \in \mathbb{R}^m$ are the process and measurement noise at time k. We assume that $w(k)$, $v(k)$, $x(0)$ are jointly independent zero mean Gaussian random variables with covariance Σ, Q and R respectively. We further assume that $Q, R \succ 0$ are strictly positive definite and (A, B) is stabilizable and (A, C) is detectable.

From system model in (1), we can write down the relation between sensor measurement y and the control input u and the noise process w and v as follows:

$$y(k) = \mathcal{G}(z)u(k) + \mathcal{H}(z)w(k) + v(k), \tag{3}$$

in which $\mathcal{G}(z) \triangleq C(zI - A)^{-1}B$ and $\mathcal{H}(z) \triangleq C(zI - A)^{-1}$, and z^{-1} is the unit delay. We assume that the controller is also a linear time invariant controller. Therefore, the control input can be written as

$$u(k) = \mathcal{K}(z)y(k). \tag{4}$$

We restrict the future discussions to the controller that satisfies the following assumption:

Assumption 1. *[Controller] The transfer function of the controller $\mathcal{K}(z)$ is a proper rational function of z. Furthermore, the closed-loop system is asymptotically stable.*

Remark 1. *If we assume that $\mathcal{K}(z)$ is rational, then $\mathcal{K}(z)$ being proper is equivalent to the controller being causal. Moreover, the limit $\lim_{z\to\infty} \mathcal{K}(\infty) < \infty$ is well-defined. For the closed-loop system, since $\mathcal{G}(z)$ is a strictly proper transfer function, it follows that $\lim_{z\to\infty} \mathcal{G}(z)\mathcal{K}(z) = 0$, which implies that $I - \mathcal{G}(z)\mathcal{K}(z)$ is invertible almost everywhere.*

We assume that an adversary passively observes the control input $u(k)$ and the sensory data $y(k)$ from time 0 to ∞. The goal of the adversary is to infer the physical system model $\mathcal{G}(z)$ from $u(k)$ and $y(k)$.

III. KPA in CPS

In this section, we shall first apply closed-loop system identification technique to the CPS and investigate the identifiability condition of $\mathcal{G}(z)$ and $\mathcal{K}(z)$ in Section III-A and then propose an algorithm to achieve so in Section III-B. A stealthy attack which is enabled by KPA is discussed in Section III-C.

A. On the identifiability of $\mathcal{G}(z)$, $\mathcal{K}(z)$

This subsection is devoted to deriving the identifiability condition of $\mathcal{G}(z)$ and $\mathcal{K}(z)$. The identifiability of such systems have been investigated based on spectral factorization.

Definition 1. *Let $e(k) = (e_1(k), .., e_N(k))^T$ be a N-dimensional discrete-time, zero-mean, wide-sense stationary random process. For any $\tau \in \mathbb{Z}$, define its autocorrelation function $R_e(\tau)$ and power spectral density $\Phi_e(z)$ as*

$$R_e(\tau) \triangleq \mathbb{E}[e(0)e^T(\tau)] = \mathbb{E}[e(k)e^T(k+\tau)].$$
$$\Phi_e(z) \triangleq \sum_{\tau=-\infty}^{\infty} R_e(\tau) z^{-\tau}.$$

Since we assume that the closed-loop system is asymptotically stable, $\begin{bmatrix} y(k) \\ u(k) \end{bmatrix}$ converges to a stationary process. Hence, the adversary can compute (or estimate) the joint power spectral density $\Phi_{y,u}$ for the limiting stationary process, if it observes the system for a sufficient amount of time. By (3) and (4), we know that $\Phi_{y,u}$ satisfies the following equation:

$$\Phi_{y,u}(z) = \mathcal{C}(z) \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \mathcal{C}^*(z). \tag{5}$$

where the closed-loop transfer function $\mathcal{C}(z)$ has the following form

$$\begin{aligned} \mathcal{C}(z) &= \begin{bmatrix} \mathcal{C}_{11}(z) & \mathcal{C}_{12}(z) \\ \mathcal{C}_{21}(z) & \mathcal{C}_{22}(z) \end{bmatrix} \\ &\triangleq \begin{bmatrix} (I-\mathcal{GK})^{-1}\mathcal{H} & (I-\mathcal{GK})^{-1} \\ \mathcal{K}(I-\mathcal{GK})^{-1}\mathcal{H} & \mathcal{K}(I-\mathcal{GK})^{-1} \end{bmatrix}. \end{aligned} \tag{6}$$

Assumption 2. *$\mathcal{C}(z)$ is asymptotically stable and minimum phase, i.e., all the poles and zeros of $\mathcal{C}(z)$ lie strictly inside the unit disk.*

Remark 2. *This is a commonly adopted assumption for input-output stability and internal stability.*

We first consider the identifiability of $\mathcal{C}(z)$ from the joint spectral density $\Phi_{y,u}$.

Lemma 1. *Under the Assumption 1 and 2, if there exists $\mathcal{C}(z)$, Q, R and $\hat{\mathcal{C}}(z)$, $\hat{Q}$, $\hat{R}$ that lead to the same $\Phi_{y,u}$, then there exists a unitary matrix V_{11}, such that*

$$\begin{aligned} \hat{\mathcal{C}}_{11}(z) &= \mathcal{C}_{11}(z)V_{11}, & \hat{\mathcal{C}}_{12}(z) &= \mathcal{C}_{12}(z), \\ \hat{\mathcal{C}}_{21}(z) &= \mathcal{C}_{21}(z)V_{11}, & \hat{\mathcal{C}}_{22}(z) &= \mathcal{C}_{22}(z), \\ \hat{Q} &= V_{11}^* Q V_{11}, & \hat{R} &= R. \end{aligned} \tag{7}$$

Proof. See Appendix. ☐

We now consider the identifiability of $\mathcal{G}(z)$, $\mathcal{K}(z)$ and $\mathcal{H}(z)$ from $\mathcal{C}(z)$. Before continuing on, we need the following definition:

Definition 2. *We define the normal rank of a transfer matrix $\mathcal{A}(z)$ to be the maximum rank of $\mathcal{A}(z)$ over all $z \in \mathbb{C}$.*

Proposition 1. *Given $\mathcal{C}(z)$, the following transfer functions can be uniquely specified :*

$$\begin{aligned} \mathcal{K}(z) &= \mathcal{C}_{22}(z)\mathcal{C}_{12}^{-1}(z), \\ \mathcal{H}(z) &= \mathcal{C}_{12}^{-1}(z)\mathcal{C}_{11}(z), \\ \mathcal{G}(z)\mathcal{K}(z) &= I - \mathcal{C}_{12}^{-1}(z). \end{aligned} \tag{8}$$

If $\mathcal{K}(z)$ has full normal row rank then $\mathcal{G}(z)$ can be uniquely determined from the following equality

$$\mathcal{G}(z) = (I - \mathcal{C}_{12}^{-1}(z))\mathcal{K}^{\dagger}(z), \tag{9}$$

where $\mathcal{K}^{\dagger}(z)$ is the unique transfer matrix satisfies $\mathcal{K}(z)\mathcal{K}^{\dagger}(z) = I$.

Proof. This is straightforward from the definition of $\mathcal{C}(z)$. ☐

Based on Lemma 1 and Proposition 1, we have the following theorem about the identifiability of $\mathcal{G}(z)$ and $\mathcal{K}(z)$.

Theorem 1. *Consider the feedback control scheme described in Sec II. Under the Assumption 1 and 2, the following statements hold:*

- *$\mathcal{G}(z)\mathcal{K}(z)$ and $\mathcal{K}(z)$ are uniquely identifiable;*
- *R is uniquely identifiable;*
- *$\mathcal{H}(z)$ and Q can be identified up to the following transformation*

$$\begin{aligned} \hat{\mathcal{H}}(z) &= \mathcal{H}(z)V_{11} \\ \hat{Q} &= V_{11}^* Q V_{11}, \end{aligned} \tag{10}$$

  in which V_{11} is a unitary matrix.

Furthermore, if $\mathcal{K}(z)$ if full normal row rank, then $\mathcal{G}(z)$ is uniquely identifiable.

Proof. Let $\mathcal{C}(z)$ be the true closed-loop transfer function. By Lemma 1, any $\hat{\mathcal{C}}(z)$ we derive from $\Phi_{y,u}$ must satisfy (7). Hence, by Proposition 1,

$$\hat{\mathcal{K}}(z) = \hat{\mathcal{C}}_{22}(z)\hat{\mathcal{C}}_{12}^{-1}(z) = \mathcal{C}_{22}(z)\mathcal{C}_{12}^{-1}(z) = \mathcal{K}(z),$$

which implies that $\mathcal{K}(z)$ is uniquely identifiable. The other statements can be proved by similar arguments. ☐

We now provide a sufficient condition under which the system is not identifiable by the adversary:

Theorem 2. *Let $w(k)$, $v(k)$ be a realization of the noise process and $x(k)$, $y(k)$, $u(k)$ be the corresponding system state, sensor measurements and control input that satisfy* (1), (2) *and* (3). *If $\mathcal{K}(z)$ can be factorized into*

$$\mathcal{K}(z) = F\tilde{\mathcal{K}}(z), \tag{11}$$

where $F \in \mathbb{R}^{p \times q}$ is a constant matrix with $q < p$ and $\tilde{\mathcal{K}}(z) \in \mathbb{C}^{q \times m}$ is a transfer function, then there exists a matrix $\hat{B} \neq B$, such that the following equalities hold for $\hat{B}$:

$$\begin{aligned} x(k+1) &= Ax(k) + \hat{B}u(k) + w(k), \\ y(k) &= Cx(k) + v(k),\ u(k) = \mathcal{K}(z)y(k). \end{aligned}$$

Proof. Since $q < p$, F is not full row rank, which implies the existence of a real matrix $\Delta B \neq 0$, such that $\Delta BF = 0$. One can verify that $\hat{B} = B + \Delta B$ is the required matrix. ☐

Remark 3. *Clearly, if the factorization described by* (11) *is possible, then the adversary cannot tell the difference between the physical system model* $\mathcal{G}(z) = C(zI - A)^{-1}B$ *and* $\hat{\mathcal{G}}(z) = C(zI - A)^{-1}\hat{B}$ *since they share the same input and output relation. This is due to the fact that the controller only inject the control input that lies in the column space of* F *and hence there are some ambiguities in the* B *matrix.*

It is worth noticing that (11) *implies that* $\mathcal{K}(z)$ *is not full normal row rank. In fact, the normal rank of* $\mathcal{K}(z)$ *is at most* q*. On the other hand, a non full normal row rank matrix* $\mathcal{K}(z)$ *can always be decomposed as* $\mathcal{K}(z) = \mathcal{F}(z)\tilde{\mathcal{K}}(z)$*, where* $\mathcal{F}(z)$ *is a* p *by* q *transfer matrix with* $q < p$*. Therefore, there exists a gap between Theorem 1 and 2. This is due to the fact that even though* $\mathcal{K}(z)$ *is not right invertible, which implies that the adversary cannot directly compute* $\mathcal{G}(z)$ *from* $\mathcal{G}(z)\mathcal{K}(z)$ *and* $\mathcal{K}(z)$*, the adversary could potentially use side information to infer* $\mathcal{G}(z)$ *(for example,* $\mathcal{G}(z) = \mathcal{H}(z)B$*.) We are planning to investigate the gap and tighten Theorem 1 and Theorem 2 in the future work.*

B. Identification algorithm

This subsection is devoted to providing a numerical algorithm for the adversary to derive $\mathcal{G}(z)$ when $\mathcal{K}(z)$ is full normal row rank, which is based on spectral factorization [19].

Since the feedback system is asymptotic stable, $\Phi_{y,u}(z)$ has no poles on the unit circle. Consider a Mobius transform $z = \frac{1+s}{1-s}$ and let $\Psi_{y,u}(s) = \Phi_{y,u}\left(\frac{1+s}{1-s}\right)$, then for $\Psi_{y,u}(s)$ there exists a positive real matrix $\mathcal{S}(s)$ [17], such that

$$\mathcal{S}(s) + \mathcal{S}^T(-s) = \Psi(s) = \mathcal{W}(s)\mathcal{W}^T(-s). \tag{12}$$

Definition 3 (Global Minimality)**.** *For a given spectral density* $\Psi(s)$*, the globally-minimal degree is the smallest degree of all its spectral factors* $\mathcal{W}(s)$*.*

Any system of globally-minimal degree is said to be *globally minimal*. Anderson [18] provides an algebraic characterization of all realizations of all spectral factors as follows. Minimal realizations of $\mathcal{S}$ are related to globally-minimal realizations of spectral factors of Ψ by the following lemma.

Lemma 2 ([18])**.** *Let* (A, B_s, C, D_s) *be a minimal realization of the positive-real matrix* $\mathcal{S}(s)$ *of* (12)*, then the system* (A, B, C, D) *is a globally-minimal realization of a spectral factor of* $\Psi(s)$*, i.e.,* $\mathcal{W}(s)$ *if and only if the following equations hold:*

$$\begin{aligned} RA^T + AR &= -BB^T \\ RC^T &= B_s - BD^T \\ 2D_s &= DD^T \end{aligned} \tag{13}$$

for some positive-definite and symmetric matrix $R \in \mathbb{R}^{n\times n}$*.*

For a properly chosen R, $\mathcal{W}(s)$ can be computed from its realization. Since $\mathcal{W}\left(\frac{z-1}{z+1}\right) = \mathcal{C}(z)D^{1/2}J \triangleq \mathcal{C}(z)\hat{D}$, for some signed identity matrix J [20]

$$\lim_{z\to\infty} \mathcal{W}\left(\frac{z-1}{z+1}\right) = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} \hat{D}. \tag{14}$$

We partition $\mathcal{W}$ and $\hat{D}$ to four blocks with corresponding dimensions as $\mathcal{C}$ in (6). Then it follows that

$$\hat{D}_{22} = \lim_{z\to\infty} \mathcal{W}_{12}\left(\frac{z-1}{z+1}\right). \tag{15}$$

Finally, once $\hat{D}_{22}$ is obtained, we can obtain an estimate closed-loop transfer function

$$\hat{\mathcal{C}}(z) = \mathcal{W}\left(\frac{z-1}{z+1}\right)\begin{bmatrix} I & 0 \\ 0 & \hat{D}_{22}^{-1} \end{bmatrix}, \tag{16}$$

and the transfer functions for plant and controller, $\mathcal{G}(z)$, $\mathcal{K}(z)$ using (8).

We summarize the identification procedure to the following Algorithm 1.

Algorithm 1 Identification algorithm for $\mathcal{G}(z)$, $\mathcal{K}(z)$

Inputs: Input-output data y_k and u_k.

Outputs: The transfer functions $\mathcal{G}$ and $\mathcal{K}$.

Step 1. *Compute* $\Phi_{y,u}(z)$ *from input-output data* y_k *and* u_k *and let* $\Psi_{y,u}(s) = \Phi_{y,u}(\frac{1+s}{1-s})$*;*

Step 2. *Each element in* $\Psi[i,j](s)$ *can be expanded as a sum of partial fractions and a term* $\Psi[i,j](\infty)$*. Those partial fractions with poles in* $Re[s] < 0$ *may then be summed together, and when add to* $\frac{1}{2}\Psi[i,j](\infty)$ *yield* $\mathcal{S}[i,j](s)$*;*

Step 3. *Compute* $\mathcal{W}(s)$ *from* (13) *for a minimal realization of* $\mathcal{S}(s)$ *and a properly chosen* $R \succeq 0$*.*

Step 4. *Compute* $\hat{D}_{22}$ *based on* $\mathcal{W}\left(\frac{z-1}{z+1}\right)$ *from* (15)*.*

Step 5. *Once an estimate* $\hat{\mathcal{C}}(z)$ *is computed using* (16) *and we can obtain* $\mathcal{G}(z)$ *and* $\mathcal{K}(z)$ *using* (8)*.*

Remark 4. *Since the main theme of this paper is to bring up the potential security issue in the classic feedback systems and propose a new control architecture which is robust to such attacks, the following numerical issues in spectral factorization are out of the scope of this paper, i..e., how the estimate of* $\Phi_{y,u}(z)$ *depends on the number of samples and how this error would propagate into the identification of* $\mathcal{G}(z)$ *and* $\mathcal{K}(z)$*.*

C. What can the attacker do after KPA?

In this section, we briefly describe a stealthy attack on the CPS after the adversary has obtained the transfer function $\mathcal{G}(z)$. The goal of this subsection is to demonstrate that KPA can enable other attacks discussed in the literature. For more detailed discuss on stealthy attack, please refer to [5].

We assume that the adversary compromised a subset of actuators and sensors and can change the corresponding control inputs and sensor measurements respectively. As a result, the system equation becomes:

$$\begin{aligned} x(k+1) &= Ax(k) + B\left[u(k) + \Gamma_u u^a(k)\right] + w(k), \\ y(k) &= Cx(k) + v(k) + \Gamma_y y^a(k), \\ u(k) &= \mathcal{K}(z)y(k), \end{aligned}$$

where $u^a(k)$ and $y^a(k)$ is the bias on the control inputs and the sensor measurements injected by the adversary at time k. Γ_u (Γ_y) is a diagonal matrix with binary diagonal elements, such that the ith diagonal elements is 1 if and only if the ith actuator (sensor) is compromised by the attacker. Since the matrices Γ_u and Γ_y represent the set of compromised actuators and sensors, they are known to the attacker. Let us define

$$\mathcal{G}_a(z) \triangleq C(zI - A)^{-1}B\Gamma_u = \mathcal{G}(z)\Gamma_u.$$

Clearly, the whole trajectory of the sensor measurements y is a function of the noise process w, v, the initial condition $x(0)$ and the adversary's action u^a, y^a. Therefore, we shall denote it as

$$y = f(w, v, x(0), u^a, y^a).$$

Notice that we omitted the control input u since u can be calculated from y.

Now if there exists a scalar $z_* \in \mathbb{C}$, and two vectors $u_* \in \mathbb{C}^p$ and $y_* \in \mathbb{C}^m$, such that

$$\mathcal{G}_a(z_*)u_* + \Gamma_y y_* = 0,$$

then the adversary can choose

$$u^a(k) = z_*^k u_*,\ y^a(k) = z_*^k y_*. \tag{17}$$

Let us define $x_* \triangleq (z_* I - A)^{-1}B\Gamma_u u_*$. One can verify that

$$f(w, v, x(0) + x_*, u^a, y^a) = f(w, v, x(0), 0, 0).$$

Therefore, the attack is stealthy since given the sensory data y, the controller cannot distinguish the following two cases from the sensory data:

1) the initial condition is $x(0) + x_*$ and the adversary injected u^a and y^a defined in (17);
2) the initial condition is $x(0)$ and no adversary exists.

Remark 5. *It is worth noticing that the adversary only need to compute z_*, u_* and y_* to launch the attack, which only requires the knowledge of $\mathcal{G}(z)$, Γ_u and Γ_y.*

IV. Low-Rank Controller Design against KPA

By Theorem 2, one way to prevent the adversary from identifying $\mathcal{G}(z)$ is to enforce the factorization (11) on the controller transfer function $\mathcal{K}(z)$. Let us define the following "virtual" control input:

$$\tilde{u}(k) \triangleq \tilde{\mathcal{K}}(z)y(k). \tag{18}$$

Hence, $u(k) = \mathcal{K}(z)y(k) = F\tilde{u}(k)$. The factorization on $\mathcal{K}(z)$ implies the CPS diagram illustrated in Fig 2.

Since we are restricting ourselves to use a low-rank controller, the performance of the system may not be optimal. In this section, we consider the problem of optimizing the following infinite horizon LQG performance:

$$J = \limsup_{T\to\infty} \frac{1}{T} \min_{u(k)} \mathbb{E}\left[\sum_{k=0}^{T-1} x(k)^T W x(k) + u(k)^T U u(k)\right], \tag{19}$$

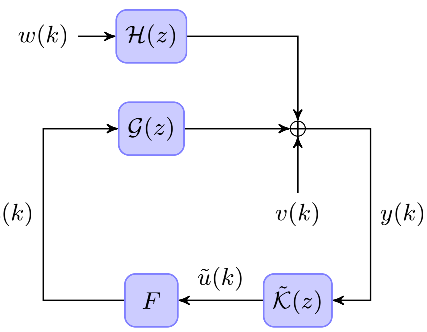

Fig. 2. The diagram of the CPS with a low-rank controller design, where $\mathcal{K}(z)$ is factorized into $F\tilde{\mathcal{K}}(z)$.

under the constraint that $F \in \mathbb{R}^{p\times q}$ where q is given. The W, U matrices are assumed to be positive semidefinite. We shall first consider how to design $\tilde{\mathcal{K}}(z)$ when F is given. We then provide a heuristic algorithm to compute the optimal F based on convex relaxation.

A. Optimal $\tilde{\mathcal{K}}(z)$

Since $u(k) = F\tilde{u}(k)$, we can rewrite the system equation as

$$x(k+1) = Ax(k) + \tilde{B}\tilde{u}(k) + w(k),$$

where $\tilde{B} \triangleq BF$. Furthermore, the objective function of LQG can be rewritten as

$$J = \limsup_{T\to\infty} \frac{1}{T} \min_{\tilde{u}(k)} \mathbb{E}\left[\sum_{k=0}^{T-1} x(k)^T W x(k) + \tilde{u}(k)^T \tilde{U} \tilde{u}(k)\right],$$

where $\tilde{U} \triangleq F^T U F \in \mathbb{R}^{q\times q}$. Therefore, the optimal control is given by a Kalman filter and a LQR controller [21]:

Kalman Filter: The state estimation of the Kalman filter (with a fixed gain) is given by:

$$\hat{x}(k) = \hat{x}(k|k-1) + K(y(k) - C\hat{x}(k|k-1)),$$
$$\hat{x}(k+1|k) = A\hat{x}(k) + Bu(k).$$

where

$$K = PC^T(CPC^T + R)^{-1},$$

and P is the fixed point of the following Riccati equation:

$$P = APA^T + Q - APC^T(CPC^T + R)^{-1}CPA^T.$$

LQR controller: The optimal control can then be derived as a linear function of the state estimate:

$$\tilde{u}(k) = \tilde{L}\hat{x}(k), \tag{20}$$

where

$$\tilde{L} = -(\tilde{B}^T\tilde{S}\tilde{B} + \tilde{U})^{-1}\tilde{B}^T\tilde{S}A,$$

and $\tilde{S}$ is the solution of the following Riccati equation

$$\tilde{S} = A^T\tilde{S}A + W - A^T\tilde{S}\tilde{B}(\tilde{B}^T\tilde{S}\tilde{B} + \tilde{U})^{-1}\tilde{B}^T\tilde{S}A. \tag{21}$$

The corresponding $\tilde{K}(z)$ is given by

$$\tilde{K}(z) = z\tilde{L}\left[zI - (I-KC)(A+B\tilde{L})\right]^{-1} K.$$

The corresponding LQG cost is given by

$$\begin{aligned} J^* &= \mathrm{tr}(\tilde{S}Q) + \mathrm{tr}[(W + A^T\tilde{S}A - \tilde{S})(P - KCP)] \\ &= \mathrm{tr}(\tilde{S}Y) + \mathrm{tr}[W(P-KCP)], \end{aligned} \tag{22}$$

where

$$\begin{aligned} Y &\triangleq Q + A(P-KCP)A^T - (P-KCP) \\ &= PC^T(CPC^T+R)^{-1}CP \succeq 0. \end{aligned} \tag{23}$$

B. Optimal F

Now we consider how to design the optimal F matrix in order to minimize the LQG cost. Since the second term on the RHS of (22) is independent of F, the optimization problem can be formulated as the following optimization problem:

$$\underset{F\in\mathbb{R}^{p\times q}}{\text{minimize}} \quad \mathrm{tr}(\tilde{S}Y). \tag{24}$$

By applying matrix inversion lemma on the RHS of (21), we have

$$\tilde{S} = A^T\left(\tilde{S}^{-1} + \tilde{B}\tilde{U}^{-1}\tilde{B}^T\right)^{-1} A + W, \tag{25}$$

where

$$\begin{aligned} &\tilde{B}\tilde{U}^{-1}\tilde{B}^T = BF\left(F^TUF\right)^{-1}F^TB \\ &= BU^{-1/2}\left[U^{1/2}F\left(F^TUF\right)^{-1}F^TU^{1/2}\right]U^{-1/2}B^T. \end{aligned}$$

Let us denote

$$X \triangleq U^{1/2}F\left(F^TUF\right)^{-1}F^TU^{1/2},\ \bar{B} \triangleq BU^{-1/2}. \tag{26}$$

It is easy to verify that $X^2 = X$ and $X = X^T$. Hence X is a symmetric projection matrix. Furthermore, $\mathrm{rank}(X) = \mathrm{rank}(F) = q$.

On the other hand, assume that X is a symmetric projection matrix of rank q. Let $v_1, \dots, v_q$ to be the orthonormal basis of the column space of X. Then the following F will satisfy (26):

$$F = U^{-1/2}\begin{bmatrix} v_1 & \dots & v_q \end{bmatrix}. \tag{27}$$

Therefore, instead of optimizing over F, the optimization problem (24) can be manipulated into

$$\begin{aligned} \underset{X\in\mathbb{S}^p}{\text{minimize}} \quad & \mathrm{tr}(\tilde{S}Y) \\ \text{subject to} \quad & \tilde{S} = g_X(\tilde{S}), \\ & X = X^T,\ X^2 = X,\ \mathrm{rank}(X) = q, \end{aligned} \tag{28}$$

where

$$g_X(\tilde{S}) \triangleq A^T\left(\tilde{S}^{-1} + \bar{B}X\bar{B}^T\right)^{-1}A + W. \tag{29}$$

We will first manipulate the constraint $\tilde{S} = g_X(\tilde{S})$ into Linear Matrix Inequalities (LMIs). To this end, we need the following intermediate result [22]:

Proposition 2. *For a fixed X, $g_X(\tilde{S})$ is monotonically non-decreasing in $\tilde{S}$.*

Consider the following optimization problem:

$$\begin{aligned} \underset{X\in\mathbb{S}^p,\,\tilde{S}}{\text{minimize}} \quad & \mathrm{tr}(\tilde{S}Y) \\ \text{subject to} \quad & \tilde{S} \geq g_X(\tilde{S}), \\ & X = X^T,\ X^2 = X,\ \mathrm{rank}(X) = q, \end{aligned} \tag{30}$$

where we relax the $\tilde{S} = g_X(\tilde{S})$ constraint in (28) to $\tilde{S} \geq g_X(\tilde{S})$. The next theorem proves that (28) and (30) are equivalent:

Lemma 3. *There exists an optimal solution $(X, \tilde{S})$ for the optimization problem* (30) *(not necessarily unique), such that the following equality holds*

$$\tilde{S} = g_X(\tilde{S}).$$

Proof. See Appendix. □

We will now rewrite the constraint $\tilde{S} \geq g_X(\tilde{S})$ as an LMI. To this end, let us take the inverse on both sides of $\tilde{S} \geq g_X(\tilde{S})$ and apply matrix inversion lemma on the RHS,

$$W^{-1} - \tilde{S}^{-1} - W^{-1}A^TZ^{-1}AW^{-1} \succeq 0, \tag{31}$$

where

$$Z = \tilde{S}^{-1} + AW^{-1}A^T + \bar{B}X\bar{B}^T.$$

Let us define $T = \tilde{S}^{-1}$, using Schur complement, we know that (31) is equivalent to:

$$\begin{bmatrix} T + AW^{-1}A^T + \bar{B}X\bar{B}^T & AW^{-1} \\ W^{-1}A^T & W^{-1} - T \end{bmatrix} \succeq 0. \tag{32}$$

Therefore, optimization problem (30) is equivalent to:

$$\begin{aligned} \underset{X,\,\tilde{S},\,T}{\text{minimize}} \quad & \mathrm{tr}(\tilde{S}Y) \\ \text{subject to} \quad & \begin{bmatrix} \tilde{S} & I \\ I & T \end{bmatrix} \succeq 0, \\ & \begin{bmatrix} T + AW^{-1}A^T + \bar{B}X\bar{B}^T & AW^{-1} \\ W^{-1}A^T & W^{-1} - T \end{bmatrix} \succeq 0, \\ & X = X^T,\ X^2 = X,\ \mathrm{rank}(X) = q. \end{aligned} \tag{33}$$

The first constraint is equivalent to $\tilde{S} \succeq T^{-1} \succeq 0$. Since we are minimizing $\mathrm{tr}(\tilde{S}Y)$ and $Y \succeq 0$, the optimal solution must have $\tilde{S} = T^{-1}$.

We will now relax the constraint on X into a convex constraint, which is given by the following lemma:

Lemma 4. *The closed convex hull of all rank q projection matrix $X \in \mathbb{S}^p$ is given by*

$$\mathbb{X} = \{X \in \mathbb{S}^p : 0 \preceq X \preceq I,\ \mathrm{tr}(X) = q\}.$$

Proof. If X is a symmetric projection matrix of rank q, then X has q eigenvalues at 1 and $p - q$ eigenvalues at 0. It is easy to verify that $\mathbb{X}$ is convex and contains X. One can further verify that the projection matrices are the extreme points of $\mathbb{X}$ and hence the lemma can be prove by Krein-Milman theorem. The detailed proof is omitted due to space limit. □

Hence, by Lemma 4, the optimization problem can be relaxed to the following semidefinite programing optimization and solved efficiently:

$$\begin{aligned} \underset{X,\tilde{S},T}{\text{minimize}} \quad & \operatorname{tr}(\tilde{S}Y) \qquad (34)\\ \text{subject to} \quad & \begin{bmatrix} \tilde{S} & I \\ I & T \end{bmatrix} \succeq 0, \\ & \begin{bmatrix} T + AW^{-1}A^T + \bar{B}X\bar{B}^T & AW^{-1} \\ W^{-1}A^T & W^{-1} - T \end{bmatrix} \succeq 0, \\ & X = X^T,\ 0 \preceq X \preceq I,\ \operatorname{tr}(X) = q. \end{aligned}$$

Remark 6. *In summary, the optimization problem* (24)*,* (28)*,* (30) *and* (33) *are all equivalent. On the other hand, the constraint on X in* (33) *is relaxed into a convex constraint in* (34)*. Therefore, the optimal value of* (34) *is no greater than the optimal value of* (24)*,* (28)*,* (30) *and* (33)*.*

Denote the optimal solution of (34) as $(X_*,\ \tilde{S}_*,\ T_*)$. Since we relaxed the constraint on X, X_* is not necessarily a projection matrix. To derive a projection matrix from X_*, one can do an eigendecomposition and rewritten X_* as

$$X_* = U_* \text{diag}(\lambda_1, \dots, \lambda_p) U_*^T,$$

where U_* is a orthonormal matrix and $\lambda_1 \geq \cdots \geq \lambda_p$. We can define a projection matrix X_0 from X_* as

$$X_0 = U_* \text{diag}(\underbrace{1, \dots, 1}_{q}, \underbrace{0, \dots, 0}_{p-q}) U_*^T.$$

Denote the corresponding fixed point of $\tilde{S} = g_{X_0}(\tilde{S})$ as $\tilde{S}_0$. Let us further denote the optimal value of (24) as α. Clearly, X_0 lies in the feasible set of the optimization problem (28). Therefore, $\operatorname{tr}(\tilde{S}_0 Y) \geq \alpha$. On the other hand, since (34) is a relaxed problem, we have $\alpha \geq \operatorname{tr}(\tilde{S}_* Y)$. Therefore, we know the optimality gap of our heuristic solution is bounded by

$$\operatorname{tr}(\tilde{S}_0 Y) - \alpha \leq \operatorname{tr}(\tilde{S}_0 Y) - \operatorname{tr}(\tilde{S}_* Y).$$

Furthermore, if X_* is indeed a projection matrix, then the optimality gap is 0 and we solve (24) exactly.

V. Numerical Example

In this section, we provide a numerical example to illustrate our low-rank controller design. We assume that $n = 15$, $m = p = 10$. The Q, R, W, U matrices are all chosen to be identity matrix. The matrix A, B, C are randomly generated, where each entry is independently drawn from a uniform distribution on $[0, 1]$.

We consider the LQG performance of a low-rank controller versus the performance of an optimal controller with no rank constraint. We define the relative performance loss as

$$\frac{J \text{ of the low rank controller}}{J \text{ of the optimal controller}} - 1.$$

We will choose q from 5 to 9 and for each q, we will perform 1000 random experiments. Fig 3 is the box and whisker diagram of the relative LQG performance loss generated by the random experiments. One can see that even if $q = 5$, meaning that we only use half degrees of freedom to design the controller, the LQG loss is still small, with median loss at 6%.

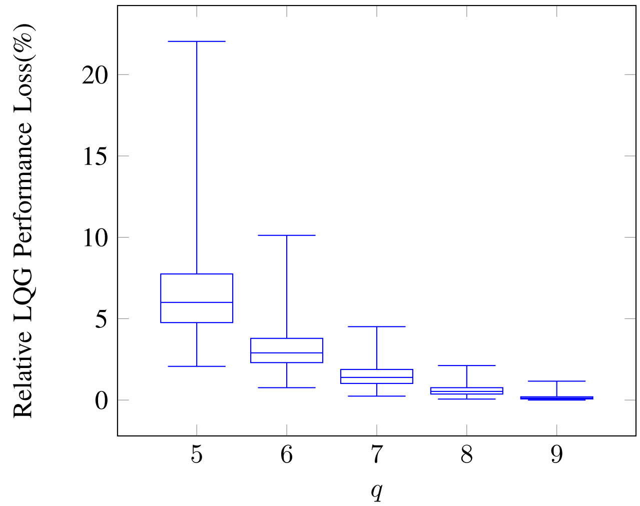

Fig. 3. Relative LQG performance Loss versus q. 1000 random experiments are conducted for each q. The bottom and top of the box represent the first and third quartiles, and the band inside the box represents the median of the data. The ends of the whiskers represent the minimum and maximum of the data.

VI. Conclusion

We consider KPA in CPS and provide a necessary condition and a sufficient condition under which the transfer function of the physical system can be uniquely identified by an adversary who passively observes the control input and sensory data. Our results demonstrate the vulnerability of the classical MIMO feedback control systems to KPA. A low-rank controller design framework is then proposed to prevent the adversary from identifying the exact physical system model. The design trade-off between system performance and security has been investigated.

VII. Acknowledgement

Ye Yuan acknowledges the support from EPSRC for supporting his visit to Murray Lab at Caltech from Dec. 2014 to Mar. 2015 during which this research has been conducted. Yilin Mo was supported in part by TerraSwarm, one of six centers of STARnet, a Semiconductor Research Corporation program sponsored by MARCO and DARPA.

The authors gratefully acknowledge Professors Richard M. Murray for the numerous interesting discussions on the topic.

References

[1] T. M. Chen, "Stuxnet, the real start of cyber warfare? [editor's note]," *IEEE Network*, vol. 24, no. 6, pp. 2–3, 2010.

[2] D. P. Fidler, "Was stuxnet an act of war? decoding a cyberattack," *IEEE Security & Privacy*, vol. 9, no. 4, pp. 56–59, 2011.

[3] Y. Liu, M. Reiter, and P. Ning, "False data injection attacks against state estimation in electric power grids," in *Proceedings of the 16th ACM conference on Computer and communications security*, 2009.

[4] S. Sundaram and C. Hadjicostis, "Distributed function calculation via linear iterative strategies in the presence of Malicious agents," *IEEE Transactions on Automatic Control*, vol. 56, no. 7, pp. 1495–1508, 2011.

[5] F. Pasqualetti, F. Dorfler, and F. Bullo, "Attack detection and identification in cyber-physical systems," *IEEE Transactions on Automatic Control*, vol. 58, no. 11, pp. 2715–2729, 2013.

[6] H. Fawzi, P. Tabuada, and S. Diggavi, "Secure estimation and control for cyber-physical systems under adversarial attacks," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1454–1467, 2014.

[7] A. Teixeira, I. Shames, H. Sandberg, and K. H. Johansson, "A secure control framework for resource-limited adversaries," *Automatica*, vol. 51, no. 0, pp. 135 – 148, 2015.

[8] Y. Mo, R. Chabukswar, and B. Sinopoli, "Detecting integrity attacks on SCada systems," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 4, pp. 1396–1407, 2014.

[9] Y. Mo, S. Weerakkody, and B. Sinopoli, "Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 93–109, 2015.

[10] R. S. Smith, "A decoupled feedback structure for covertly appropriating networked control systems," 2011, pp. 90–95.

[11] A. Teixeira, K. C. Sou, H. Sandberg, and K. Johansson, "Secure control systems: A quantitative risk management approach," *IEEE Control Systems Magazine*, vol. 35, no. 1, pp. 24–45, 2015.

[12] K. Koscher, A. Czeskis, F. Roesner, S. Patel, T. Kohno, S. Checkoway, D. McCoy, B. Kantor, D. Anderson, H. Shacham, and S. Savage, "Experimental security analysis of a modern automobile," in *Security and Privacy (SP), 2010 IEEE Symposium on*, 2010, pp. 447–462.

[13] T. Ng, G. Goodwin, and B. Anderson, "Identifiability of mimo linear dynamic systems operating in closed loop," *Automatica*, vol. 13, no. 5, pp. 477–485, 1977.

[14] B. Anderson, "An algebraic solution to the spectral factorization problem," *IEEE Transactions on Automatic Control*, vol. 12, no. 4, pp. 410–414, 1967.

[15] B. Anderson and M. Gevers, "Identifiability of linear stochastic systems operating under linear feedback," *Automatica*, vol. 18, no. 2, pp. 195–213, 1982.

[16] L. Ljung, *System identification*. Springer, 1998.

[17] K. Glover, "Structural aspects of system identification," 1973.

[18] B. D. Anderson, "The inverse problem of stationary covariance generation," *Journal of Statistical Physics*, vol. 1, no. 1, pp. 133–147 0022–4715, 1969.

[19] A. H. Sayed and T. Kailath, "A survey of spectral factorization methods," *Numerical linear algebra with applications*, vol. 8, no. 67, pp. 467–496, 2001.

[20] D. Hayden, Y. Yuan, and J. Goncalves, "Network reconstruction from intrinsic noise," *arXiv preprint arXiv:1310.0375*, 2013.

[21] L. Schenato, B. Sinopoli, M. Franceschetti, K. Poolla, and S. S. Sastry, "Foundations of Control and Estimation Over Lossy Networks," *Proc. IEEE*, vol. 95, no. 1, pp. 163–187, 2007.

[22] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, and S. S. Sastry, "Kalman filtering with intermittent observations," *Automatic Control, IEEE Transactions on*, vol. 49, no. 9, pp. 1453–1464, 2004.

APPENDIX

Proof of Lemma 1. From the definition, $\Phi_{y,u}(w)$ is real rational and positive semi-definite for $|z| = 1$. The closed-loop transfer function $\mathcal{C}(z)$ is stable and minimum phase. Therefore, $\mathcal{C}(z)$ is analytic in $|z| \geq 1$. Since $\mathcal{G}, \mathcal{H}$ are strictly proper, we have $\mathcal{G}(\infty) = 0$, $\mathcal{H}(\infty) = 0$. On the other hand, since $\mathcal{K}$ is proper and rational, $\mathcal{K}(\infty)$ exists. Hence

$$\lim_{z\to\infty} \mathcal{C}(z) = \begin{bmatrix} 0 & I \\ 0 & \mathcal{K}(\infty) \end{bmatrix}.$$

Assume that both $\mathcal{C},\ D = \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix}$ and $\hat{\mathcal{C}},\ \hat{D} = \begin{bmatrix} \hat{Q} & 0 \\ 0 & \hat{R} \end{bmatrix}$ give the same $\Phi_{y,u}$ satisfying

1) D and $\hat{D}$ are block diagonal and positive definite matrices;
2) both $\mathcal{C}$ and $\hat{\mathcal{C}}$ are stable and minimum phase,

then there exists a paraunitary matrix $\mathcal{V}(z)$ such that [15]

$$\hat{\mathcal{C}}(z) = \mathcal{C}(z)\mathcal{V}(z), \tag{35}$$

$$\hat{D} = \mathcal{V}(z)D\mathcal{V}^*(z). \tag{36}$$

From (35), since both $\mathcal{C}(z)$ and $\hat{\mathcal{C}}(z)$ are stable and minimum phase, $\mathcal{V}(z)$ is stable and minimum phase, which implies that $\mathcal{V}(z)$ is a constant matrix independent of z [18], [20]. Therefore, we denote it simply as V. Take $z \to \infty$ on both sides of (35) yields

$$\begin{bmatrix} 0 & I \\ 0 & \hat{\mathcal{K}}(\infty) \end{bmatrix} = \begin{bmatrix} 0 & I \\ 0 & \mathcal{K}(\infty) \end{bmatrix} \lim_{z\to\infty} V, \tag{37}$$

which leads to

$$V_{21} = 0,\ V_{22} = I. \tag{38}$$

Since $VV^* = I$, we have $V_{12} = 0$ and $V_{11}V_{11}^* = I$. As a result, (35) and (36) imply that

$$\begin{aligned} &\hat{\mathcal{C}} = \mathcal{C}\begin{bmatrix} V_{11} & 0 \\ 0 & I \end{bmatrix} \Leftrightarrow \begin{cases} \hat{\mathcal{C}}_{11} = \mathcal{C}_{11}V_{11} \\ \hat{\mathcal{C}}_{12} = \mathcal{C}_{12} \\ \hat{\mathcal{C}}_{21} = \mathcal{C}_{21}V_{11} \\ \hat{\mathcal{C}}_{22} = \mathcal{C}_{22} \end{cases} \quad \text{and} \\ &\hat{D} = \begin{bmatrix} V_{11}^* & 0 \\ 0 & I \end{bmatrix} D \begin{bmatrix} V_{11} & 0 \\ 0 & I \end{bmatrix} \Leftrightarrow \begin{cases} \hat{Q} = V_{11}^* Q V_{11} \\ \hat{R} = R \end{cases}. \end{aligned} \tag{39}$$

□

Proof to Lemma 3. Assume that $(X, \tilde{S})$ is the optimal solution for (30). Since $\tilde{S} \geq g_X(\tilde{S})$ and g_X is monotonically non-decreasing in $\tilde{S}$, we know that

$$\tilde{S} \geq g_X(\tilde{S}) \geq g_X^{(2)}(\tilde{S}) \geq \cdots \geq 0,$$

where

$$g_X^{(1)}(\tilde{S}) \triangleq g_X(\tilde{S}),\ g_X^{(n+1)}(\tilde{S}) \triangleq g_X\left(g_X^{(n)}(\tilde{S})\right).$$

Since $g_X^{(n)}(\tilde{S})$ is monotonically decreasing and positive semidefinite, it will converge to a matrix $\tilde{S}^* = g_X(\tilde{S}^*) \leq \tilde{S}$. Therefore, $(X, \tilde{S}^*)$ is also the optimal solution of (30), which finishes the proof. □