2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Safety-Preserving Filters Against Stealthy Sensor and Actuator Attacks

Cédric Escudero[1], Carlos Murguia[2], Paolo Massioni[3] and Eric Zamaï[4]

Abstract— **This article proposes a novel strategy based on control input filtering for mitigating the effects of deception attacks on control and sensor measurement signals. We assume an adversary who can tamper data transmitted from a communication network in order to degrade the plant performance. The proposed strategy consists on adding multiple-input multiple-output (MIMO) filters to the control loop, between the received control actions and the plant actuators. The filter's goal is to dynamically steer the reachable set induced by the attack signals to a safe region of the state space. The article provides a filter synthesis method under the form of a semidefinite programming problem, yielding such filters in a way that attack-free control signals are distorted as little as possible, and plant trajectories are contained in the safe set. At the end of the paper, a set of simulations demonstrate the effectiveness of the approach.**

I. INTRODUCTION

The connectivity of control systems is rapidly growing, with the goal of increasing performance. On the other hand, this often means that controllers are now connected with the plant though a network that might be unsecured, increasing the risk of cyberattacks [1]. In particular, attacks that tamper with system's signals (sensing and control) are now considered as one of the main threat for control systems. Such attacks aim to disrupt the control system's operations in order to degrade its performance, by means of the so-called *deception* attacks [2]. The control engineering community is actively investigating in order to cope with such attacks [1], [3]. A special attention is directed towards stealthy attacks, a type of attack that aims to remain invisible with respect to all safety-enforcing agents acting on the closed-loop system (e.g. fault detectors, operators) [4]–[6].

The literature features a number of promising recently-developed methods to reduce the effect of (stealthy) attacks on closed-loop systems. These methods aim to design/redesign elements of the closed-loop system (e.g. controller, fault detector) to mitigate the effect of (stealthy) attacks [3], and they are often referred to as mitigation methods. Among the mitigation methods, set-theoretic tools relying on the computation of reachable sets have revealed their potential to make sure that the plant states avoid dangerous states. This last approach includes attack analysis to quantify the effect of (stealthy) attacks against systems [7], and then safety enforcement to force the system trajectories to avoid physical degradation [8]–[11]. It includes design of secure controllers [8]–[10], design of secure fault detector [7], [12] and saturation of control signals [11].

In this article, we consider False Data Injection (FDI) attacks, a class of deception attacks that inject malicious signals into the true sensing and control signals. We assume these attacks aim to drive the plant dynamics to a part of the state space leading to physical degradation. We propose a set-theoretic method to synthesize a filter placed between the plant and a remote controller to guarantee the plant safety. In a previous work [13], a method has been presented to synthesize filters for open-loop systems (the plant) that might be subject to actuator attacks only. Such filters process the control inputs before they reach the plant, enforcing by design that actuator attacks cannot lead the plant to physical degradation. Here, we propose a set-theoretic method to synthesize safety-preserving filters for closed-loop systems, including a plant, a controller, and a fault detector. Moreover, we now consider stealthy actuator and sensor attacks that try to avoid raising an alarm in the fault detector. To the best of our knowledge, stealthy actuator attacks with respect to a fault detector have not yet been explored in set-theoretic methods.

The remainder of this manuscript is organized as follows. Section II presents the system under study and states the research problem. Section III provides a preliminary tool (reachability and ellipsoidal approximations of reachable sets) to perform the filter synthesis. Our main results to synthesize the filters are presented in Section IV. Lastly in Section V, we apply our results on a simulation example to illustrate the performance of our tools.

Notation: The symbol $\mathbb{R}$ stands for the real numbers, $\mathbb{R}^{n \times m}$ is the set of real $n \times m$ matrices, and $\mathbb{R}_{>0}$ ($\mathbb{R}_{\geq 0}$) denotes the set of positive (non-negative) real numbers. Matrix $A^\top$ indicates the transpose of matrix A and $\text{diag}(a_1, ..., a_n)$ corresponds to a diagonal matrix with diagonal elements $a_1, ..., a_n$. The identity matrix of dimension n is denoted by I_n, and $\mathbf{0}$ is a matrix of only zeros of appropriate dimensions. The notation $A \succeq 0$ (resp. $A \preceq 0$) indicates that the matrix A is positive (resp. negative) semidefinite, i.e., all the eigenvalues of the symmetric matrix A are positive (resp. negative) or equal to zero, whereas the notation $A \succ 0$ (resp. $A \prec 0$) indicates the positive (resp. negative) definiteness, i.e., all the eigenvalues are strictly positive (resp. negative). The notation $\mathcal{E}(\Phi, \bar{\phi})$ stands for an ellipsoidal set of dimension φ with

[1] Cédric Escudero is with Univ Lyon, INSA Lyon, Université Claude Bernard Lyon 1, Ecole Centrale de Lyon, CNRS, Ampère, UMR5005, 69621 Villeurbanne, France cedric.escudero@insa-lyon.fr

[2] Carlos Murguia is with the Department of Mechanical Engineering, Dynamics and Control Group, Eindhoven University of Technology, The Netherlands c.g.murguia@tue.nl

[3] Paolo Massioni is with Univ Lyon, INSA Lyon, Université Claude Bernard Lyon 1, Ecole Centrale de Lyon, CNRS, Ampère, UMR5005, 69621 Villeurbanne, France paolo.massioni@insa-lyon.fr

[4] Eric Zamaï is with Univ Lyon, INSA Lyon, Université Claude Bernard Lyon 1, Ecole Centrale de Lyon, CNRS, Ampère, UMR5005, 69621 Villeurbanne, France eric.zamai@insa-lyon.fr

979-8-3503-0123-6/23/$31.00 ©2023 IEEE

shape matrix $\Phi \in \mathbb{R}^{\varphi \times \varphi}$, $\Phi \succ 0$ and centered at $\bar{\phi}$. For ellipsoids centered at the origin, we simply write $\mathcal{E}(\Phi)$.

II. System Description and Problem Statement

In this section, we present the class of systems and attacks under study, and state the research problem.

A. Plant Dynamics (Σ_p)

We consider linear time-invariant plants Σ_p of the form:

$$\Sigma_p \begin{cases} \dot{x}_p(t) = A_p x_p(t) + B_p u_p(t), \\ y_p(t) = C_p x_p(t), \end{cases} \tag{1}$$

with time $t \in \mathbb{R}_{>0}$, plant state $x_p(t) \in \mathbb{R}^{n_p}$, control input $u_p(t) \in \mathbb{R}^m$, sensor measurements $y_p(t) \in \mathbb{R}^l$, and plant matrices $A_p \in \mathbb{R}^{n_p \times n_p}$, $B_p \in \mathbb{R}^{n_p \times m}$ and $C_p \in \mathbb{R}^{l \times n_p}$ with stabilizable (A_p, B_p) and detectable (A_p, C_p). The plant transmits $y_p(t)$ through an unsecured/public network to a remote station equipped with a dynamic controller Σ_c and an anomaly detector Σ_d, see Figure 1. The remote station receives a networked version, $\tilde{y}(t)$, of the system output $y_p(t)$. Vector $\tilde{y}(t)$ is used by Σ_d to compute alarm signals and by Σ_c to compute control actions $u_c(t)$, which are sent back to the plant. The plant receives a networked version, $\tilde{u}(t)$, of u_c to close the loop. Both $\tilde{y}(t)$ and $\tilde{u}(t)$ are subject to potential FDI attacks, $\delta_y(t)$ and $\delta_u(t)$, at the network, see Figure 1.

B. Controller Dynamics (Σ_c)

We consider output dynamic controllers Σ_c of the form:

$$\Sigma_c \begin{cases} \dot{x}_c(t) = A_c x_c(t) + B_c \tilde{y}(t), \\ u_c(t) = C_c x_c(t) + D_c \tilde{y}(t), \end{cases} \tag{2}$$

with controller state $x_c(t) \in \mathbb{R}^{n_c}$, networked sensor data $\tilde{y}(t) \in \mathbb{R}^l$, control actions $u_c(t) \in \mathbb{R}^m$, and matrices $A_c \in \mathbb{R}^{n_c \times n_c}$, $B_c \in \mathbb{R}^{n_c \times m}$, $C_c \in \mathbb{R}^{m \times n_c}$ and $D_c \in \mathbb{R}^{m \times l}$. We assume that, for $\tilde{y}(t) = y_p(t)$, i.e., no network effects, the plant in closed loop with controller Σ_c has a globally asymptotically stable equilibrium point.

C. Adversarial Capabilities

We consider FDI attacks that aim to drive the plant dynamics to a part of the state space where physical degradation occurs – referred here to as critical states. These critical states could model, for instance, the inter-vehicle distance that should satisfy safety constraints. We assume the adversary is capable of additively injecting signals, $\delta_u(t)$ and $\delta_y(t)$, to true control actions u_c and/or true sensor measurements $y_p(t)$, respectively, at the unsecured network. The adversary can compromise up to s_u control actions, $s_u \in \{1, \ldots, m\}$, and s_y sensor measurements, $s_y = \{1, \ldots, l\}$. We introduce adversary's selection matrices, Λ_u and Λ_y, to be able to select how the additive signals $\delta_u(t)$ and $\delta_y(t)$ affect $u_c(t)$ and $y_p(t)$ (for sensitivity analysis). Hence, the transmitted control actions, $\tilde{u}(t)$, and the transmitted sensor measurements, $\tilde{y}(t)$ take the form:

$$\begin{cases} \tilde{u}(t) = u_c(t) + \Lambda_u \delta_u(t), \\ \tilde{y}(t) = y_p(t) + \Lambda_y \delta_y(t), \end{cases} \tag{3}$$

with additive actuator attack $\delta_u(t) \in \mathbb{R}^{s_u}$, additive sensor attack $\delta_y(t) \in \mathbb{R}^{s_y}$, and adversary's selection matrices $\Lambda_u \in \mathbb{R}^{m \times s_u}$, and $\Lambda_y \in \mathbb{R}^{l \times s_y}$ We assume resource-limited adversaries that can inject bounded signals $\delta_u(t)$ and $\delta_y(t)$. We remark that most FDI attacks have constraints in the signals that they can inject due to physical limitations (power/bandwidth), computing limits (speed, memory), and/or attack strategy (stealthiness, jamming, replay). We capture limited resources as hard ellipsoidal bounds, $\mathcal{E}_u(\mathcal{U}, \bar{u})$ and $\mathcal{E}_y(\mathcal{Y}, \bar{y})$, on the injected signals $(\delta_u(t), \delta_y(t))$:

$$\mathcal{E}_u(\mathcal{U}, \bar{u}) := \left\{ \delta_u | (\delta_u - \bar{u})^\top \mathcal{U} (\delta_u - \bar{u}) \leqslant 1 \right\}, \tag{4}$$

$$\mathcal{E}_y(\mathcal{Y}, \bar{y}) := \left\{ \delta_y | (\delta_y - \bar{y})^\top \mathcal{Y} (\delta_y - \bar{y}) \leqslant 1 \right\}, \tag{5}$$

for some known positive definite matrices $\mathcal{U} \in \mathbb{R}^{s_u \times s_u}$, $\mathcal{Y} \in \mathbb{R}^{s_y \times s_y}$ and vectors $\bar{u} \in \mathbb{R}^{s_u}$, $\bar{y} \in \mathbb{R}^{s_y}$.

D. Fault Detector (Σ_d)

We consider a remote station equipped with a model-based fault detector, Σ_d, to pinpoint the occurrence of faults and attacks in the plant dynamics. We consider residual-based detectors comprised of a Luenberger state observer and a static change detection rule:

$$\Sigma_d \begin{cases} \dot{\hat{x}}_p(t) = A_p \hat{x}_p(t) + B_p u_c(t) + L r(t), \\ r(t) = \tilde{y}(t) - C_p \hat{x}_p(t), \\ a(t) = \begin{cases} 1 & \text{if } (r(t) - \bar{r})^\top \Pi (r(t) - \bar{r}) > 1, \\ 0 & \text{if } (r(t) - \bar{r})^\top \Pi (r(t) - \bar{r}) \leq 1, \end{cases} \end{cases} \tag{6}$$

with estimated plant state $\hat{x}_p(t) \in \mathbb{R}^{n_p}$, alarm signal $a(t) \in \mathbb{R}$, observer gain $L \in \mathbb{R}^{n_p \times l}$, positive definite detector matrix $\Pi \in \mathbb{R}^{l \times l}$, $\Pi \succ 0$, and detector center $\bar{r} \in \mathbb{R}^l$. We assume the observer gain L and detector parameters $(\Pi, \bar{r})$ have been designed such that $(A_p - LC_p)$ is Hurwitz and $a(t) = 0$ in the absence of faults/attacks. That is, a successful detection occurs when $a(t) = 1$ and either $\delta_u(t)$ or $\delta_y(t)$ are different from zero for some $t \geq 0$. Note that, given (1), (3), and (6), the estimation error $e(t) := x_p(t) - \hat{x}_p(t)$ satisfies the following differential equation:

$$\Sigma_e \left\{ \dot{e}(t) = (A_p - LC_p) e - L \Lambda_y \delta_y(t) + B_p \Lambda_u \delta_u(t) \right\}. \tag{7}$$

Hence, the estimation error $e(t)$ and residual $r(t)$ converge to the origin in the attack-free case.

E. Attack Models

We consider two types of FDI attacks, *stealthy* and *non-stealthy*. *Stealthy attacks* aim to damage the integrity of the system while enforcing the fault detector in (6) not to raise alarms. Stealthy attacks are constrained in the sense that the adversary must carefully choose ($\delta_u(t)$, $\delta_y(t)$) such that $(r(t) - \bar{r})^\top \Pi (r(t) - \bar{r}) \leq 1$. They are usually slow persistent attacks with low power. On the other hand, *(non-stealthy) attacks* aim to damage the integrity of the plant while ignoring the detector. Non-stealthy attacks often induce fast and/or large damage to the system, but have an increased risk of being detected. Let $\mathcal{E}_r(\Pi, \bar{r})$ denote the stealthy set of residuals:

$$\mathcal{E}_r(\Pi, \bar{r}) := \left\{ r | (r - \bar{r})^\top \Pi (r - \bar{r}) \leq 1 \right\}, \tag{8}$$

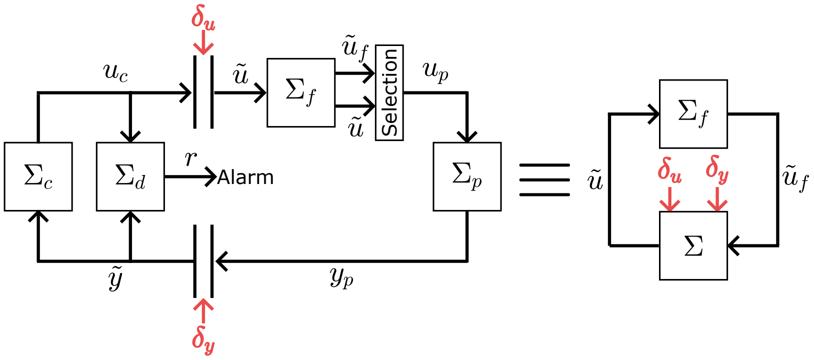

Fig. 1: Closed-loop system under attacks with the safety-preserving filter to synthesize.

If the injected ($\delta_u(t)$, $\delta_y(t)$) enforce that the residual satisfies (8), no alarms will be raised in the fault detector (so we say that these attacks are stealthy). In this article, we cover both stealthy and non-stealthy attacks. Results vary slightly from one case to the other.

To characterize the safety in the presence of attacks, we introduce the following notion of *safe sets*.

Definition 1 (Safe Set) *The safe set $X_s \subseteq \mathbb{R}^n$ of plant Σ_p in* (1) *is the set of states $x_p \in X_s$ that guarantee a safe and proper operation of the plant. The safe set X_s is the part of the state space that excludes critical states – states that, if reached, compromise the plant physical integrity.*

Safe sets exclude, by definition, all critical states from the state space of the plant in (1).

In the remainder of this section, we present the filter dynamics that we aim to synthesize, and state the research problem we seek to address.

F. Safety-Preserving Filters (Σ_f)

We propose to protect the plant Σ_p in (1) against (stealthy) actuator/sensor attacks by filtering control actions, $\tilde{u}(t)$, that might have been corrupted by attack signals $(\delta_u(t), \delta_y(t))$ in (3) before they are applied to the plant. That is, we pass $\tilde{u}(t)$ through a filter to enforce, by design, that it is impossible for (stealthy) actuator/sensor attacks to drive the plant outside the safe set X_s. The filter output, $\tilde{u}_f(t)$, reaches the plant to shape the closed-loop system reachable set, see Figure 1. We consider linear time-invariant filters of the form:

$$\Sigma_f \begin{cases} \dot{x}_f(t) = A_f x_f(t) + B_f \tilde{u}(t), \\ \tilde{u}_f(t) = C_f x_f(t) + D_f \tilde{u}(t), \end{cases} \tag{9}$$

with filter state $x_f(t) \in \mathbb{R}^{n_f}$, filter input $\tilde{u}(t) \in \mathbb{R}^m$ (potentially corrupted control actions received from the network), filter output $\tilde{u}_f(t) \in \mathbb{R}^m$ (the control inputs to be applied to the plant), and filter matrices $A_f \in \mathbb{R}^{n_f \times n_f}$, $B_f \in \mathbb{R}^{n_f \times m}$, $C_f \in \mathbb{R}^{m \times n_f}$, and $D_f \in \mathbb{R}^{m \times m}$ to be designed. We allow for partial filtering in the sense that not all control actions $\tilde{u}(t)$ are filtered. That is, we allow for some $\tilde{u}(t)$ to get through the filter and reach the system directly (without having been distorted by the filter). It follows that the control inputs driving the plant, $u_p(t)$, can be written as follows:

$$u_p(t) = \Gamma_c \tilde{u}(t) + \Gamma_f \tilde{u}_f(t), \tag{10}$$

where $\Gamma_c \in \mathbb{R}^{m \times m}$ and $\Gamma_f \in \mathbb{R}^{m \times m}$ are diagonal selection matrices used to select which control actions are unfiltered/filtered. Matrices Γ_c and Γ_f satisfy:

$$\Gamma_c + \Gamma_f = I_m. \tag{11}$$

G. Closed-Loop System Dynamics (Σ)

After having defined the plant dynamics (1), controller dynamics (2), attack signals (3), and fault detector dynamics (6), we can now write the stacked closed-loop dynamics, Σ, under FDI attacks as follows:

$$\Sigma \begin{cases} \dot{z}(t) = Az(t) + B\tilde{u}_f(t) + E\delta_y(t) + F\delta_u(t), \\ \tilde{u}(t) = Cz(t) + G\delta_y(t) + H\delta_u(t), \end{cases} \tag{12}$$

with stacked state $z := [x_p^\top, x_c^\top, e^\top]^\top \in \mathbb{R}^{n_z}$, $n_z = 2n_p + n_c$, and matrices A, B, E, F, C, G, and H given in (13).

We next write the dynamics Σ in (13) in feedback interconnection with the filter in (9) (see Figure 1). Define the extended state $\zeta := [z^\top, x_f^\top]^\top \in \mathbb{R}^n$, $n = n_z + n_f$. Then, the filtered system under attacks can be writte as follows:

$$\dot{\zeta}(t) = \tilde{A}\zeta(t) + \tilde{E}\delta_y(t) + \tilde{F}\delta_u(t). \tag{14}$$

with matrices $\tilde{A}$, $\tilde{E}$, and $\tilde{F}$ given in (13).

We can now state the research problem we seek to address.

Problem 1 *Given the stacked dynamics* (12)*, the filter dynamics* (9)*, the safe set X_s in Definition 1, and the stealthy set $\mathcal{E}_r$ in* (8)*, find filter matrices (A_f,B_f,C_f,D_f) such that the plant trajectories are contained in X_s for all resource-limited actuator/sensor injection attacks satisfying* (4)-(5).

The solution to Problem 1 aims to enforce that the state trajectories of (1) subject to (stealthy) actuator/sensor attacks, in series interconnection with the filter (9), are constrained inside the safe set X_s.

III. Preliminary Results

In this section, we introduce some preliminary results from [14] that will be used to derive the main result of the manuscript (the solution to Problem 1).

$$A := \begin{bmatrix} A_p + B_p\Gamma_c D_c C_p & B_p\Gamma_c C_c & \mathbf{0} \\ B_c C_p & A_c & \mathbf{0} \\ B_p\Gamma_c D_c C_p - B_p D_c C_p & B_p\Gamma_c C_c - B_p C_c & A_p - LC_p \end{bmatrix}, B := \begin{bmatrix} B_p\Gamma_f \\ \mathbf{0} \\ B_p\Gamma_f \end{bmatrix}, E := \begin{bmatrix} B_p\Gamma_c D_c\Lambda_y \\ B_c\Lambda_y \\ B_p\Gamma_c D_c\Lambda_y - B_p D_c\Lambda_y - L\Lambda_y \end{bmatrix},$$
$$F := \begin{bmatrix} B_p\Gamma_c\Lambda_u \\ \mathbf{0} \\ B_p\Gamma_c\Lambda_u \end{bmatrix}, C := \begin{bmatrix} D_c C_p & C_c & \mathbf{0} \end{bmatrix}, \tilde{A} := \begin{bmatrix} A + BD_f C & BC_f \\ B_f C & A_f \end{bmatrix}, \tilde{E} := \begin{bmatrix} BD_f G + E \\ B_f G \end{bmatrix}, \tilde{F} := \begin{bmatrix} BD_f H + F \\ B_f H \end{bmatrix},$$
$$G := \begin{bmatrix} D_c\Lambda_y \end{bmatrix}, H := \begin{bmatrix} \Lambda_u \end{bmatrix}, \tag{13}$$

Definition 2 (Reachable Set) *The reachable set $\mathcal{R}_\zeta(t)$ at time $t \in \mathbb{R}_{>0}$ from initial condition $\zeta(t_0) \in \mathbb{R}^n$ is the set of states $\zeta(t)$ that satisfy the differential equation* (14)*, over all attack signals $(\delta_u(t), \delta_y(t))$ satisfying* (4)-(5)*, i.e.,*

$$\mathcal{R}_\zeta(t) := \left\{ \zeta(t) \left| \begin{array}{l} \zeta(t_0) \in \mathbb{R}^n, \\ \zeta(t) \text{ satisfies } (14),\ \delta_u(t) \in \mathcal{E}_u(\mathcal{U}, \bar{u}), \\ \text{and } \delta_y(t) \in \mathcal{E}_y(\mathcal{Y}, \bar{y}). \end{array} \right. \right\}. \tag{15}$$

Note that, because the attack signals $(\delta_u(t), \delta_y(t))$ are bounded, the set $\mathcal{R}_\zeta(t)$ always exists if A in (14) is Hurwitz; which is true when filter and system matrices A_f, A are both Hurwitz due to the block triangular structure of $\tilde{A}$.

A. Ellipsoidal Outer Approximation of $\mathcal{R}_\zeta(t)$

Because the exact computation of $\mathcal{R}_\zeta(t)$ is not tractable, we compute an outer ellipsoidal approximation $\mathcal{E}_\zeta(Q)$ of $\mathcal{R}_\zeta(t)$ using a set-theoretic method reliying on some properties of positively invariant sets [15].

Definition 3 *The ellipsoidal set $\mathcal{E}_\zeta(Q)$ is invariant for the dynamical system* (14)*, if for all initial states $\zeta(t_0) \in \mathcal{E}_\zeta(Q)$, and all $\delta_u(t) \in \mathcal{E}_u(\mathcal{U}, \bar{u})$, $\delta_y(t) \in \mathcal{E}_y(\mathcal{Y}, \bar{y})$, the trajectories $\zeta(t)$ of* (14) *satisfy $\zeta(t) \in \mathcal{E}_\zeta(Q), \forall\, t \geq 0$.*

By Definition 3, any invariant ellipsoidal set $\mathcal{E}_\zeta(Q)$ is an outer ellipsoidal approximation of $\mathcal{R}_\zeta(t)$. In previous work [14], we have provided sufficient conditions for ellipsoidal sets to be invariant for a class of linear-time invariant systems as (12) subject to peak-bounded inputs as $\delta_u(t)$ and $\delta_y(t)$ and when $\zeta(t)$ is constrained to remain inside a given set $\mathcal{E}(\Xi, \bar{\xi})$ with $\Xi \succeq 0$. The method is based on the search of a Lyapunov-like function, $V(\zeta) = \zeta^\top Q\zeta$, with certain properties using Linear Matrix Inequalities (LMIs) [16]. Here, instead of having $\zeta(t) \in \mathcal{E}(\Xi, \bar{\xi})$, we want to constrain the residuals $r(t)$ to belong to the stealthy set $\mathcal{E}_r$ (for stealthy attacks), i.e., $r(t) \in \mathcal{E}_r(\Pi, \bar{r})$. Hence, the constraint changes. First, from the definition of $r(t)$ in (6), it is easy to verify that $r(t) = C_p e(t) + \Lambda_y\delta_y(t)$, where $e(t)$ is the estimation error satisfying (7). Then, the stealthy set formulation (8) can be written as an inequality in terms of estimation error $e(t)$ and the attack signal $\delta_y(t)$. Moreover, because $e(t)$ can be written in terms of the extended state $\zeta(t)$ as $e(t) = \Gamma\zeta(t)$ with $\Gamma = [\mathbf{0}\ \mathbf{0}\ I_{n_p}\ \mathbf{0}]$, the stealthy set (8) can be written in terms of $\zeta(t)$. We can now state the following lemma, adapted from [14], used to find invariant ellipsoidal sets for the closed-loop system dynamics (14) with peak-bounded attack signals $(\delta_u(t), \delta_y(t))$ (3) and residuals $r(t)$ within the stealthy set $\mathcal{E}_r(\Pi, \bar{r})$.

Lemma 1 (Invariant Ellipsoidal Set [14]) *Consider the extended system dynamics* (13)-(14)*. If there exist matrix $Q \in \mathbb{R}^{n\times n}$ and constants α, β, λ, $\rho \in \mathbb{R}_{\geq 0}$ satisfying:*

$$-J - \alpha K - \beta L - \lambda M - \rho N \succeq 0, \tag{16}$$
$$Q \succ 0, \tag{17}$$

with

$$J = \begin{bmatrix} \tilde{A}^\top Q + Q\tilde{A} & \mathbf{0} & Q\tilde{E} & Q\tilde{F} \\ * & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ * & * & \mathbf{0} & \mathbf{0} \\ * & * & * & \mathbf{0} \end{bmatrix},$$
$$K = \begin{bmatrix} Q & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ * & -1 & \mathbf{0} & \mathbf{0} \\ * & * & \mathbf{0} & \mathbf{0} \\ * & * & * & \mathbf{0} \end{bmatrix}, L = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ * & 1 - \bar{y}^\top\mathcal{Y}\bar{y} & \bar{y}^\top\mathcal{Y} & \mathbf{0} \\ * & * & -\mathcal{Y} & \mathbf{0} \\ * & * & * & \mathbf{0} \end{bmatrix},$$
$$M = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ * & 1 - \bar{u}^\top\mathcal{U}\bar{u} & \mathbf{0} & \bar{u}^\top\mathcal{U} \\ * & * & \mathbf{0} & \mathbf{0} \\ * & * & * & -\mathcal{U} \end{bmatrix},$$
$$N = \begin{bmatrix} -\Gamma^\top C_p^\top\Pi C_p\Gamma & \Gamma^\top C_p^\top\Pi\bar{r} & -\Gamma^\top C_p^\top\Pi\Lambda_y & \mathbf{0} \\ * & 1 - \bar{r}^\top\Pi\bar{r} & \bar{r}^\top\Pi\Lambda_y & \mathbf{0} \\ * & * & -\Lambda_y^\top\Pi\Lambda_y & \mathbf{0} \\ * & * & * & \mathbf{0} \end{bmatrix}; \tag{18}$$

then, $\zeta(0)^\top Q\zeta(0) \leqslant 1 \Rightarrow \zeta(t)^\top Q\zeta(t) \leqslant 1$, for all $t \geq 0$, $\delta_y(t) \in \mathcal{E}_y(\mathcal{Y}, \bar{y})$, $\delta_u(t) \in \mathcal{E}_u(\mathcal{U}, \bar{u})$, and $r(t) \in \mathcal{E}_r(\Pi, \bar{r})$.

To consider the case of non-stealthy attacks, we just remove the term ρN from (16) in Lemma 1. This term provides the extra constraint needed only if the adversary tries to avoid raising an alarm in the fault detector.

IV. Problem Formulation and Solution

In this section, we propose a synthesis framework, built around Lemma 1, to find filter matrices solving Problem 1 in terms of the solution of a series of semidefinite programs.

A. Safety Enforcement

To prevent damage from (stealthy) attacks, the plant states must remain inside the safe set X_s. Here, we model/embed X_s as an ellipsoid $\mathcal{E}_s(\Psi, \bar{\psi})$ satisfying (19), which can be written in terms of the stacked system state $z(t)$:

$$\mathcal{E}_s(\Psi, \bar{\psi}) := \{z | (z - \bar{\psi})^\top\Psi(z - \bar{\psi}) \leqslant 1\}, \tag{19}$$

with a known positive semi-definite matrix $\Psi \in \mathbb{R}^{n_z \times n_z}$ and vector $\bar{\psi} \in \mathbb{R}^{n_z}$. Matrix Ψ could be, in general, rank-deficient, as only part of the plant states might be subject to safety constraints – $\mathcal{E}_s$ can even coincide with $\mathbb{R}^{n_z \times n_z}$ by picking $\Psi = \mathbf{0}$, meaning that none of the system states are subject to a safe zone.

To enforce safety, we want to guarantee that the system states z belong to the safe set $\mathcal{E}_s(\Psi, \bar{\psi})$. If the conditions of Lemma 1 are satisfied, all trajectories of the extended state $\zeta = [z^\top, x_f^\top]^\top$ belong to the ellipsoidal set $\mathcal{E}_\zeta(Q)$. Because we want to enforce safety on the system states $z(t)$ (see (19)), we work with the projection of $\mathcal{E}_\zeta(Q)$ onto the z-hyperplane (the system state space). By [17, Appendix A.3, Lemma 10], this projection is also an ellipsoid, $\mathcal{E}_z(Q_z)$, with shape matrix $Q_z = Q_1 - Q_2 Q_3^{-1} Q_2^\top$ and

$$Q := \begin{bmatrix} Q_1 & Q_2 \\ * & Q_3 \end{bmatrix}. \tag{20}$$

Hence, if the conditions of Lemma 1 hold, $z(t) \in \mathcal{E}_z(Q_z)$, and therefore, to guarantee safety, we require

$$z(t) \in \mathcal{E}_z(Q_z) \subseteq \mathcal{E}_s(\Psi, \bar{\psi}). \tag{21}$$

B. Distortion Constraint

By filtering $\tilde{u}(t)$, we degrade the control performance as we are changing the dynamics of the control signals applied to the plant. To reduce this degradation, we introduce a distortion metric that quantifies the difference between $\tilde{u}(t)$ and $u_p(t)$ in the frequency domain. Define the distortion signal $w(t) := u_p(t) - \tilde{u}(t)$ in the attack-free case, i.e., for $\delta_y(t) = \delta_u(t) = \mathbf{0}$. By replacing $u_p(t)$ by (10), $\tilde{u}$ by (12) and $\tilde{u}_f$ by (9), the distortion signal $w(t)$ can be written in terms of the extended state $\zeta(t)$ as $w(t) = C_w \zeta(t)$ with

$$C_w := \begin{bmatrix} \Gamma_c C + \Gamma_f D_f C - C & \Gamma_f C_f \end{bmatrix} \tag{22}$$

We treat this $w(t)$ as a performance output for the closed-loop dynamics (14). Note that, in the filter-free case, $u_p(t) = \tilde{u}(t)$, so $w(t) = \mathbf{0}$ for all $t \geq t_0$. To reduce the degradation due to the filter, we want to make $z(t)$ small in some appropriate sense. For system (14), with input $\tilde{u}(t)$ and output $w(t)$, let $T_{\tilde{u} \to w}(s)$ denote the transfer matrix from $\tilde{u}(t)$ to $w(t)$, i.e., $T_{\tilde{u} \to w}(s) := C_w(sI_n - \tilde{A})^{-1}$. Given this transfer matrix, we use its H_∞ norm to quantify the effect of $\tilde{u}(t)$ on $w(t)$, i.e., $||T_{\tilde{u} \to w}(s)||_{H_\infty}$. If no filter is in place, the norm is trivially zero, and as we let $\tilde{u}$ and u_p be more different, the norm grow unbounded. An upper bound on this norm is used to shape the filter dynamics so that the change in the dynamics of control inputs is constraint. That is, when designing the filter to guarantee safety, we also seek to enforce that the norm of $T_{\tilde{u} \to w}(s)$ is below a predefined level $\gamma \in \mathbb{R}_{\geq 0}$. We use this γ to modulate how much we are willing to sacrifice in terms of control performance to enforce safety. By the bounded-real lemma [16], $||T_{\tilde{u} \to w}(s)||_{H_\infty}$ is less than or equal to $\gamma \in \mathbb{R}_{\geq 0}$, if there exists a positive definite matrix Q and constant $\epsilon \in \mathbb{R}_{\geq 0}$ satisfying:

$$L_{H_\infty} := \begin{bmatrix} \tilde{A}^\top Q + Q\tilde{A} & \mathbf{0} & C_w^\top \\ * & -(\gamma - \epsilon) I_m & \mathbf{0} \\ * & * & -\gamma I_m \end{bmatrix} \preceq 0 \tag{23}$$

C. Filter Synthesis Problem

After having derived conditions to (i) synthesize the filter so that safety of the plant is guaranteed; and (ii) limit the change of dynamics in the control inputs, we can now re-cast Problem 1 above in terms of our new notation.

Problem 2 (Filter Synthesis Problem) *Find the filter matrices $\kappa := (A_f,\ B_f,\ C_f,\ D_f)$ such that (i) the ellipsoidal set $\mathcal{E}_\zeta(Q)$ is invariant under the closed-loop system dynamics* (14) *for attack signals $\delta_u(t) \in \mathcal{E}_u(\mathcal{U}, \bar{u})$, $\delta_y(t) \in \mathcal{E}_y(\mathcal{Y}, \bar{y})$, and residuals $r(t) \in \mathcal{E}_r(\Pi, \bar{r})$ (for stealthy attacks); (ii) the system state $z(t)$ belong to the safe set $\mathcal{E}_s(\Psi, \bar{\psi})$, i.e., $z(t) \in \mathcal{E}_s(\Psi, \bar{\psi})$; and (iii) the $||T_{\tilde{u} \to w}(s)||_{H_\infty}$ is upper bounded by γ.*

An optimal solution to Problem 2 can be obtained by solving an optimization problem; however, because now $\kappa := (A_f,\ B_f,\ C_f,\ D_f)$ are variables in the synthesis problem, the blocks $Q\tilde{A}$, $Q\tilde{E}$ and $Q\tilde{F}$ in matrix J of the constraint (16) are not convex in (κ,Q). Following the results in [18], we propose an invertible linearizing change of variables such that, in the new variables, we can cast an equivalent synthesis program that has convex cost and affine constraints.

D. Change of Variables and Convex Reformulation

Let Q be positive definite and of the form:

$$Q = \begin{bmatrix} Y & N \\ N^\top & \tilde{Y} \end{bmatrix},\ Q^{-1} = \begin{bmatrix} X & M \\ M^\top & \tilde{X} \end{bmatrix}, \tag{24}$$

where Y, N, $\tilde{Y}$, X, M, $\tilde{X} \in \mathbb{R}^{n \times n}$; and Y, $\tilde{Y}$, X, $\tilde{X}$ are positive definite matrices. Define the following matrices

$$\Pi_1 := \begin{bmatrix} X & I_n \\ M^\top & \mathbf{0} \end{bmatrix},\ \Pi_2 := \begin{bmatrix} I_n & Y \\ \mathbf{0} & N^\top \end{bmatrix}. \tag{25}$$

Using block matrix inversion formulas, it is easy to verify that $YX + NM^\top = I$ and $N^\top X + \tilde{Y} M^\top = \mathbf{0}$, and therefore $Q\Pi_1 = \Pi_2$. Define the change of filter variables:

$$\begin{cases} \hat{A}_f := NA_fM^\top + NB_fCX + YBC_fM^\top + YAX + YBD_fCX, \\ \hat{B}_f := NB_f + YBD_f, \\ \hat{C}_f := C_fM^\top + D_fCX, \\ \hat{D}_f := D_f, \end{cases} \tag{26}$$

with $\hat{A}_f \in \mathbb{R}^{n \times n}$, $\hat{B}_f \in \mathbb{R}^{n \times m}$, $\hat{C}_f \in \mathbb{R}^{m \times n}$, $\hat{D}_f \in \mathbb{R}^{m \times m}$. If M and N have full rank, and the synthesis variables $\nu := (X, Y, \hat{A}_f, \hat{B}_f, \hat{C}_f, \hat{D}_f)$ are given, we can always extract filter matrices (A_f, B_f, C_f, D_f) satisfying (26). The aim is to formulate an optimization problem that is convex in the synthesis variables ν.

Invariance: First consider the invariance constraints, (16)-(17). Blocks $Q\tilde{A}$, $Q\tilde{E}$ and $Q\tilde{F}$ in matrix J are now nonlinear in (κ,Q). The aim is to find equivalent constraints that are affine in ν. Define $\mathcal{Q}(\nu) := \Pi_1^\top Q \Pi_1$ with Π_1 defined in

(25). Using matrix inversion formulas, we have $\mathcal{Q}(\nu) = \mathrm{diag}(X, Y)$, which is linear in ν. If Π_1 is invertible, $\mathcal{Q}(\nu)$ is a congruence transformation of Q; and therefore $Q \succ 0 \Leftrightarrow \mathcal{Q}(\nu) \succ 0$, [16]. Moreover, by Schur complement properties, we have $\mathcal{Q}(\nu) \succ 0 \Leftrightarrow (X \succ 0$ and $YX - I \succ 0)$. Because we have $YX + NM^\top = I$ (see the text below (25)), we can conclude that $\mathcal{Q}(\nu) \succ 0$ implies $YX - I = -NM^\top \succ 0$. Then, using singular value decomposition of $YX - I$, we can always find full rank N and M, which implies that Π_1 is invertible for $X \succ 0$. Hence, the following matrix inequality will replace (17) in the synthesis problem:

$$\mathcal{Q}(\nu) = \begin{bmatrix} X & I_n \\ I_n & Y \end{bmatrix} \succ 0. \tag{27}$$

Next, consider (16) and define $\mathcal{F} := \mathrm{diag}[\Pi_1, I, I]$. $\mathcal{F}$ is invertible if $\mathcal{Q}(\nu) \succ 0$ (see the above discussion). It follows that, for invertible $\mathcal{F}$, (17) is equivalent to $\mathcal{F}^\top\big(-J - \alpha K - \beta L - \lambda M - \rho N\big)\mathcal{F} \succeq 0$. Define the matrices:

$$J' := \mathcal{F}^\top J \mathcal{F} = \begin{bmatrix} \mathcal{A}(\nu)^\top + \mathcal{A}(\nu) & \mathbf{0} & \mathscr{E}(\nu) & \mathcal{F}(\nu) \\ * & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ * & * & \mathbf{0} & \mathbf{0} \\ * & * & * & \mathbf{0} \end{bmatrix}, \tag{28}$$

$$K' := \mathcal{F}^\top K \mathcal{F} = \begin{bmatrix} \mathcal{Q}(\nu) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ * & -1 & \mathbf{0} & \mathbf{0} \\ * & * & \mathbf{0} & \mathbf{0} \\ * & * & * & \mathbf{0} \end{bmatrix}, \tag{29}$$

$$L' := \mathcal{F}^\top L \mathcal{F} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ * & 1 - \bar{y}^\top \mathcal{Y} \bar{y} & \bar{y}^\top \mathcal{Y} & \mathbf{0} \\ * & * & -\mathcal{Y} & \mathbf{0} \\ * & * & * & \mathbf{0} \end{bmatrix}, \tag{30}$$

$$M' := \mathcal{F}^\top M \mathcal{F} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ * & 1 - \bar{u}^\top \mathcal{U} \bar{u} & \mathbf{0} & \bar{u}^\top \mathcal{U} \\ * & * & \mathbf{0} & \mathbf{0} \\ * & * & * & -\mathcal{U} \end{bmatrix}, \tag{31}$$

$$N' := \begin{bmatrix} -\mathcal{G}(\nu) & \mathcal{H}(\nu) & -\mathcal{I}(\nu) & \mathbf{0} \\ * & 1 - \bar{r}^\top \Pi \bar{r} & \bar{r}^\top \Pi \Lambda_y & \mathbf{0} \\ * & * & -\Lambda_y^\top \Pi \Lambda_y & \mathbf{0} \\ * & * & * & \mathbf{0} \end{bmatrix} \succeq \mathcal{F}^\top N \mathcal{F}, \tag{32}$$

where

$$\begin{aligned}
\mathcal{A}(\nu) &:= \Pi_1^\top Q \tilde{A} \Pi_1 = \begin{bmatrix} AX + B\hat{C}_f & A + B\hat{D}_f C \\ \hat{A}_f & YA + \hat{B}_f C \end{bmatrix}, \\
\mathscr{E}(\nu) &:= \Pi_1^\top Q \tilde{E} = \begin{bmatrix} B\hat{D}_f G + E \\ \hat{B}_f G + YE \end{bmatrix}, \\
\mathcal{F}(\nu) &:= \Pi_1^\top Q \tilde{F} = \begin{bmatrix} B\hat{D}_f H + F \\ \hat{B}_f H + YF \end{bmatrix}, \\
\mathcal{G}(\nu) &:= \begin{bmatrix} \bar{\mathcal{R}} X + X \bar{\mathcal{R}}^\top - I_n & X\mathcal{R}' \\ * & \mathcal{R}' \end{bmatrix} \preceq \Pi_1^\top (\Gamma^\top C_p \Pi C_p \Gamma) \Pi_1, \\
\mathcal{I}(\nu) &:= \Pi_1^\top \Gamma^\top C_p^\top \Pi \Lambda_y = \begin{bmatrix} X\mathcal{R}''' \\ \mathcal{R}''' \end{bmatrix}, \\
\mathcal{H}(\nu) &:= \Pi_1^\top \Gamma^\top C_p^\top \Pi \bar{r} = \begin{bmatrix} X\mathcal{R}'' \\ \mathcal{R}'' \end{bmatrix}, \\
\mathcal{R}' &:= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ * & C_p^\top \Pi C_p \end{bmatrix}, \mathcal{R}'' := \begin{bmatrix} \mathbf{0} \\ C_p^\top \Pi \bar{r} \end{bmatrix}, \mathcal{R}''' := \begin{bmatrix} \mathbf{0} \\ C_p^\top \Pi \Lambda_y \end{bmatrix},
\end{aligned} \tag{33}$$

where $\mathcal{R}' =: \bar{\mathcal{R}}\bar{\mathcal{R}}^\top$ using Cholesky decomposition of $\mathcal{R}'$.

Note that (28)-(32) coming from the change of coordinates in (26) are linear in ν. We have now $\mathcal{F}^\top\big(-J - \alpha K - \beta L - \lambda M - \rho N\big)\mathcal{F} \succeq -J' - \alpha K' - \beta L' - \lambda M' - \rho N'$ with N' a linear in ν upper bound on $\mathcal{F}^\top N \mathcal{F}$, i.e., (17) is satisfied if

$$-J' - \alpha K' - \beta L' - \lambda M' - \rho N' \succeq 0. \tag{34}$$

Because (34) is linear in ν, we will replace (17) with the sufficient condition (34) in the synthesis problem.

Safety: Using the partition of Q in (24), the change of coordinates in (26), and the projection lemma in [17, Appendix A.3, Lemma 10], we can verify that $\mathcal{E}_z(X^{-1}) := \{z | z^\top X^{-1} z \leq 1\}$, with X from (24). Recall that, for safety we require $\mathcal{E}_z(X^{-1}) \subseteq \mathcal{E}_s(\Psi, \bar{\psi})$. By writting these ellipsoids as quadratic inequalities, it follows from the S-procedure that $\mathcal{E}_z(X^{-1}) \subseteq \mathcal{E}_s(\Psi, \bar{\psi})$ if and only if there exists $\delta \in \mathbb{R}_{\geq 0}$ satisfying $J_1 - \delta W_1 \preceq 0$ with

$$J_1 := \begin{bmatrix} \Psi & -\Psi\bar{\Psi} \\ * & \bar{\Psi}^\top \Psi \bar{\Psi} - 1 \end{bmatrix}, W_1 := \begin{bmatrix} X^{-1} & \mathbf{0} \\ * & -1 \end{bmatrix}.$$

Because W_1 is written in terms of X^{-1}, W_1 is nonlinear in ν. To tackle this, we can perform a congruence transformation $\mathcal{F}_1 := \mathrm{diag}[X, 1]$. Recall that X is positive definite from the invariance constraint, so $\mathcal{F}_1$ is invertible. It follows that $\mathcal{F}_1^\top (J_1 - \delta W_1) \mathcal{F}_1 \preceq 0$ is a congruence transformation of the safety inequality $J_1 - \delta W_1$. Then, using properties of Schur complements, we can conclude that $\mathcal{F}_1^\top (J_1 - \delta W_1) \mathcal{F}_1 \preceq 0$ if and only if the following matrix inequality is satisfied

$$\begin{bmatrix} \delta X & X\Psi\bar{\Psi} & -X \\ * & -\bar{\Psi}^\top \Psi \bar{\Psi} + 1 - \delta & \mathbf{0} \\ * & * & \Psi^{-1} \end{bmatrix} \succeq 0. \tag{35}$$

Because (35) is affine in ν, we will use it as a safety constraint in the synthesis problem.

Distortion: Consider (23) and $\mathcal{F} = \mathrm{diag}[\Pi_1, I, I]$, with $\mathcal{F}$ invertible as $\mathcal{Q}(\nu) \succ 0$ (see above). Then, (23) is satisfied if and only if $L'_{H_\infty} := \mathcal{F}^\top L_{H_\infty} \mathcal{F} \preceq 0$, which is linear in ν :

$$L'_{H_\infty} = \begin{bmatrix} \mathcal{A}(\nu)^\top + \mathcal{A}(\nu) & \mathbf{0} & \mathcal{C}_w(\nu)^\top \\ * & -(\gamma - \epsilon) I_m & \mathbf{0} \\ * & * & -\gamma I_m \end{bmatrix} \preceq 0. \tag{36}$$

with $(\mathcal{A}(\nu), \mathcal{B}(\nu))$ as defined in (33), and $\mathcal{C}_w(\nu)$ given by:

$$\mathcal{C}_w(\nu) := C_w \Pi_1 = \begin{bmatrix} \Gamma_f \hat{C}_f^\top + \Gamma_c C X - C X & \Gamma_c C + \Gamma_f \hat{D}_f C - C \end{bmatrix}.$$

Hence, (23) is replaced by the constraint (36) in the synthesis program. We can now state the main result of this article.

Theorem 1 *Consider the closed-loop system* (14) *with matrices in* (13)*, the stealthy set* $\mathcal{E}_r(\Pi, \bar{r})$ *in* (8)*, and the safe set* $\mathcal{E}_s(\Psi, \bar{\psi})$ *in* (19)*. For given filtering selection matrices* Γ_c*,* Γ_f *satisfying* (11)*, if there exist* $X \in \mathbb{R}^{n \times n}$*,* $Y \in \mathbb{R}^{n \times n}$*,* $n = n_z + n_f$*,* $\hat{A}_f \in \mathbb{R}^{n_z \times n_z}$*,* $\hat{B}_f \in \mathbb{R}^{n_z \times m}$*,* $\hat{C}_f \in \mathbb{R}^{m \times n_z}$*,* $\hat{D}_f \in \mathbb{R}^{m \times m}$*, and* α*,* β*,* λ*,* ρ*,* δ*,* γ*,* $\epsilon \in \mathbb{R}_{\geq 0}$ *satisfying* (27)*,* (34)*,* (35)*, and* (36)*; then, for all* $t \geq t_0$*,* $\delta_u(t) \in \mathcal{E}_u(\mathcal{U}, \bar{u})$*,* $\delta_y(t) \in \mathcal{E}_y(\mathcal{Y}, \bar{y})$*, and* $r(t) \in \mathcal{E}_r(\Pi, \bar{r})$*, we have:* **(a)** $\zeta(0)^\top Q \zeta(0) \leqslant 1 \Rightarrow \zeta(t)^\top Q \zeta(t) \leqslant 1$*;* **(b)** $z(t) \in \mathcal{E}_z(X^{-1}) \subseteq \mathcal{E}_s(\Psi, \bar{\psi})$*; and* **(c)** $||T_{\tilde{u} \to w}(s)||_{H_\infty} \leq \gamma$*.*

Proof: Theorem 1 follows from Lemma 1 and the discussion in Section IV.

Remark 1 (Non-Stealthy Case) *To consider the case of non-stealthy attacks, we just need to remove* $\rho N'$ *from* (34) *in Theorem 1 (see text below Lemma 1).*

Theorem 1 provides sufficient conditions to synthesize a safety-preserving filter ($\hat{A}_f$, $\hat{B}_f$, $\hat{C}_f$, $\hat{D}_f$) that guarantees the system state trajectories $z(t)$, including the plant state trajectories $x_p(t)$, to remain inside the safe set $\mathcal{E}_s$, and the distortion metric is below a predefined level γ. Here, we aim to design filters that lead to the smallest invariant ellipsoidal set of system states $\mathcal{E}_z$. As mentioned before, $\mathcal{E}_z$ can be written as $\mathcal{E}_z(X^{-1}) := \{z | z^\top X^{-1} z \leq 1\}$, with X from (24). As a criterion for the ellipsoid size, we use the trace of the shape matrix P for an ellipsoid $\mathcal{E}(P)$ (see [16] for details). Hence to reduce the size of $\mathcal{E}_z(X^{-1})$ we minimize the trace of X in the synthesis program.

Filter Synthesis

$$
\mathbf{OP}_1\!: \begin{cases} \min\limits_{\nu} \ \text{trace}(X), & \textit{(minimum volume } \mathcal{E}_z(X^{-1})\textit{)} \\ \text{s.t. } (27), (34); & \textit{(invariance)} \\ \quad (35); & \textit{(safety)} \\ \quad (36). & \textit{(distortion)} \end{cases}
$$

Due to products of α with $\mathcal{Q}(\nu)$, ρ with $\mathcal{G}$, $\mathcal{H}$, and δ with X, some matrix inequalities in Theorem 1 are quasi-convex, i.e., for fixed α, ρ, δ the constraint is a LMI. To relax it, we solve $\mathbf{OP}_1$ repeatedly for different values of $\alpha, \rho, \delta \geq 0$ until the smallest cost is attained. By solving $\mathbf{OP}_1$, optimal X, Y, and ($\hat{A}_f$, $\hat{B}_f$, $\hat{C}_f$, $\hat{D}_f$) are obtained. These matrices are then used to extract the filter matrices $\kappa := (A_f, B_f, C_f, D_f)$ following the procedure in [18]. Firstly, we compute M and N satisfying $MN^\top = I_n - XY$. Secondly, we compute Π_1 and Π_2 in (25) and extract $Q = \Pi_2 \Pi_1^{-1}$. Lastly, we extract filter matrices $\kappa := (A_f, B_f, C_f, D_f)$ by sequentially solving (26) for D_f, C_f, B_f and A_f in this order.

V. Simulation Example

Consider a system as in (12) with a plant Σ_p as in (1) with system matrices in (37) defined for $x_p = [x_{p1}, x_{p2}, x_{p3}]^\top$ ($n_p = 3$), $u_p = [u_{p1}, u_{p2}]^\top$ ($m = 2$), and $y_p = [y_{p1}, y_{p2}, y_{p3}]^\top$ ($l = 3$) where $C_p = I_l$; a controller Σ_c as in (2) defined as $u_c(t) = D_c \tilde{y}(t)$ ($n_c = 0$); a fault detector Σ_d that raises an alarm for $\Pi = \text{diag}(10, 10, 10)$, $\bar{r} = \mathbf{0}$.

$$
A_p = \begin{bmatrix} -10 & 10 & 10 \\ 0 & -150 & 0 \\ 0 & 0 & -150 \end{bmatrix}, \ B_p = \begin{bmatrix} 0 & 0 \\ 100 & 1 \\ 0 & 100 \end{bmatrix},
$$

$$
D_c = \begin{bmatrix} 0 & 1.41 & 0.01 \\ 0 & 0.01 & 1.41 \end{bmatrix}, \ L = \begin{bmatrix} 70 & 10 & 10 \\ 0 & -50 & 0 \\ 0 & 0 & -50 \end{bmatrix}. \quad (37)
$$

For the safe set $\mathcal{E}_s(\Psi, \bar{\psi})$, we use $\bar{\psi} = \mathbf{0}$ and shape matrix $\Psi = \text{diag}(0.05, 0.05, 0.05, 1 \times 10^{-4}, 1 \times 10^{-4}, 1 \times 10^{-4})$. Consider actuator and sensor attacks (δ_u, δ_y) satisfying (4), and (5) with $\mathcal{U} = \text{diag}(10, 10)$, $\bar{u} = \mathbf{0}$, $\mathcal{Y} = \text{diag}(10, 10, 10)$, $\bar{y} = \mathbf{0}$.

First, we analyze the effect of the attacks on the closed-loop system (when no filter is placed), i.e., $u_p(t) = \tilde{u}(t)$ (see Figure 1). We use Lemma 1 and find the optimal Q that minimizes the size of the ellipsoidal set $\mathcal{E}_\zeta(Q)$ with $\alpha = 4$, $\rho = 1$, $\epsilon = 10^{-8}$. The result is drawn in Figure 2 where the projection of the ellipsoidal sets $\mathcal{E}_\zeta$ onto the z-hyperplane ($\mathcal{E}_z$) and onto the r-hyperplane ($\mathcal{E}_r$) are the ellipsoids in dotted curves. The safe set $\mathcal{E}_s$ and the stealthy set $\mathcal{E}_r$ are respectively the blue filled ellipsoid and the ellipsoid with red cross markers. It is easy to observe that the ellipsoid $\mathcal{E}_z$ is not a subset of the safe set $\mathcal{E}_s$. This means stealthy attacks can drive the system trajectories outside the safe set.

To force that the plant trajectories are within the safe set in the presence of attacks, we now synthesize a safety preserving filter acting on the control signals $\tilde{u}$ as in (9) before they reach the plant. The filter we seek to synthesize acts on the complete $\tilde{u}(t)$, i.e., $\Gamma_c = \mathbf{0}$, $\Gamma_f = I_m$. We solve the optimization problem $\mathbf{OP}_1$ with $\alpha = 2$, $\rho = 1$, $\delta = 0.5$, $\epsilon = 10^{-8}$, $\gamma = 5.5 \times 10^{-6}$. The results are shown in Figure 2, where the projection of the ellipsoidal set $\mathcal{E}_\zeta$ onto the z-hyperplane ($\mathcal{E}_z$) and the r-hyperplane ($\mathcal{E}_r$) are the ellipsoids filled in green and cyan, respectively. Note that the filter manages to push the reachable set within the safe, i.e., $\mathcal{E}_z \subseteq \mathcal{E}_s$. Hence, the effect of stealthy actuator and sensor attacks is mitigated by placing the computed filter in the loop. In Figure 3, the Bode diagram of the computed filter is shown. We can see that at low frequency, the filter attenuates slightly the control inputs, while the attenuation is amplified in higher frequency.

VI. Conclusion

We have derived a set-theoretic method to synthesize optimal LTI filters that constrain control inputs to avoid reachability of unsafe/critical states induced by resource-limited (stealthy) actuator/sensor attacks. The filter synthesis is posed as the solution of a convex optimization problem where we constrain how much we are willing to sacrifice in term of control performance to enforce safety. The use of these filters allows constraining control inputs dynamically as we can impose amplification limits in the frequency domain. While the method is presented for control input filtering only, it can also be applied to design filters for sensor

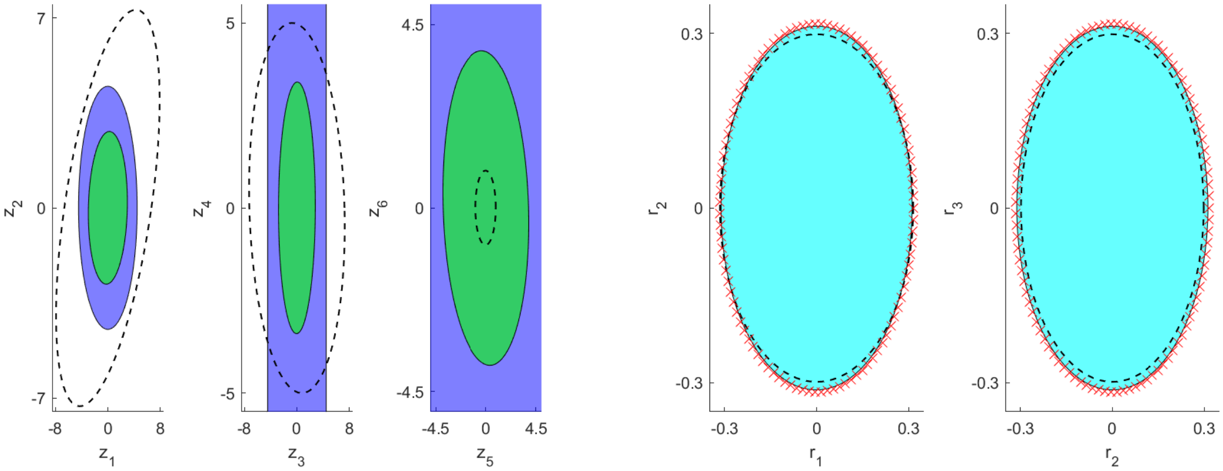

Fig. 2: On the left: projection of the invariant ellipsoidal set $\mathcal{E}_\zeta(Q)$ onto the z-hyperplane (without the filter: dotted line ; with the filter: green fill), safe set $\mathcal{E}_s(\Psi, \bar{\psi})$ (blue fill). On the right: projection of the invariant ellipsoidal set $\mathcal{E}_\zeta(Q)$ onto the r-hyperplane (without the filter: dotted line ; with the filter: cyan fill), stealthy set $\mathcal{E}_r(\Pi, \bar{r})$ (red cross marker).

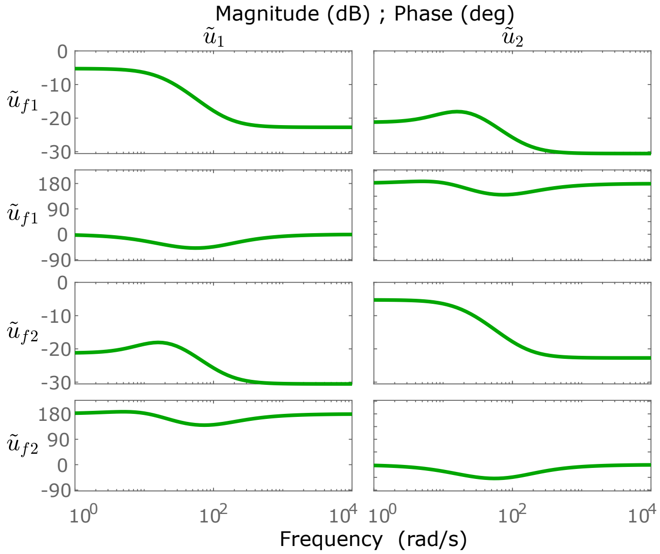

Fig. 3: Bode diagram of the safety-preserving filter Σ_f

measurements. However, the co-design of both filters is yet not feasible due to the high order of the resulting filters. This will be explored in future work.

Acknowledgement

The research leading to these results has received funding from the European Union's Horizon Europe programme under grant agreement No 101069748 – SELFY project.

References

[1] S. M. Dibaji, M. Pirani, D. B. Flamholz, A. M. Annaswamy, K. H. Johansson, and A. Chakrabortty, "A systems and control perspective of cps security," *Annu. Rev. Control*, vol. 47, pp. 394 – 411, 2019.

[2] A. Teixeira, D. Pérez, H. Sandberg, and K. H. Johansson, "Attack models and scenarios for networked control systems," in *Proceedings of the 1st International Conference on High Confidence Networked Systems*, ser. HiCoNS '12. New York, NY, USA: ACM, 2012.

[3] J. Giraldo, D. Urbina, A. Cardenas, J. Valente, M. Faisal, J. Ruths, N. O. Tippenhauer, H. Sandberg, and R. Candell, "A survey of physics-based attack detection in cyber-physical systems," *ACM Comput. Surv.*, vol. 51, no. 4, pp. 1–36, Jul. 2018.

[4] A. Zaman, B. Safarinejadian, and W. Birk, "Security analysis and fault detection against stealthy replay attacks," *International Journal of Control*, vol. 95, no. 6, pp. 1562–1575, 2022.

[5] S. C. Anand and A. M. H. Teixeira, "Stealthy cyber-attack design using dynamic programming," in *2021 60th IEEE Conference on Decision and Control (CDC)*, 2021, pp. 3474–3479.

[6] Q. Zhang, K. Liu, A. M. H. Teixeira, Y. Li, S. Chai, and Y. Xia, "An online kullback-leibler divergence-based stealthy attack against cyber-physical systems," *IEEE Transactions on Automatic Control*, pp. 1–8, 2022.

[7] Y. Mo and B. Sinopoli, "On the performance degradation of cyber-physical systems under stealthy integrity attacks," *IEEE Transactions on Automatic Control*, vol. 61, no. 9, pp. 2618–2624, 2016.

[8] S. C. Anand and A. M. H. Teixeira, "Risk-averse controller design against data injection attacks on actuators for uncertain control systems," in *2022 American Control Conference (ACC)*, 2022.

[9] Y. Lin, M. S. Chong, and C. Murguia, "Plug-and-play secondary control for safety of LTI systems under attacks," *CoRR*, vol. abs/2212.00593, 2022.

[10] K. Gheitasi and W. Lucia, "A worst-case approach to safety and reference tracking for cyber-physical systems under network attacks," *IEEE Transactions on Automatic Control*, pp. 1–7, 2022.

[11] S. Hadizadeh Kafash, N. Hashemi, C. Murguia, and J. Ruths, "Constraining attackers and enabling operators via actuation limits," in *2018 IEEE Conference on Decision and Control (CDC)*, 2018.

[12] C. Murguia and J. Ruths, "On model-based detectors for linear time-invariant stochastic systems under sensor attacks," *IET Control Theory Appl.*, vol. 13, no. 8, pp. 1051–1061, May 2019.

[13] C. Escudero, C. Murguia, P. Massioni, and E. Zamaï, "Enforcing safety under actuator injection attacks through input filtering," in *2022 European Control Conference (ECC)*, 2022, pp. 1521–1528.

[14] C. Escudero, P. Massioni, E. Zamai, and B. Raison, "Analysis, prevention, and feasibility assessment of stealthy ageing attacks on dynamical systems," *IET Control Theory Appl.*, July 2021.

[15] F. Blanchini, "Set invariance in control," *Automatica*, vol. 35, no. 11, pp. 1747 – 1767, 1999.

[16] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory*. SIAM, 1994, vol. 15.

[17] C. Murguia, I. Shames, J. Ruths, and D. Nešić, "Security metrics and synthesis of secure control systems," *Automatica*, vol. 115, May 2020.

[18] C. Scherer, P. Gahinet, and M. Chilali, "Multiobjective output-feedback control via lmi optimization," *IEEE Transactions on Automatic Control*, vol. 42, no. 7, pp. 896–911, 1997.