A TRACTABLE FAULT DETECTION AND ISOLATION APPROACH FOR NONLINEAR SYSTEMS WITH PROBABILISTIC PERFORMANCE

PEYMAN MOHAJERIN ESFAHANI AND JOHN LYGEROS

ABSTRACT. This article presents a novel perspective along with a scalable methodology to design a fault detection and isolation (FDI) filter for high dimensional nonlinear systems. Previous approaches on FDI problems are either confined to linear systems or they are only applicable to low dimensional dynamics with specific structures. In contrast, shifting attention from the system dynamics to the disturbance inputs, we propose a relaxed design perspective to train a linear residual generator given some statistical information about the disturbance patterns. That is, we propose an optimization-based approach to robustify the filter with respect to finitely many signatures of the nonlinearity. We then invoke recent results in randomized optimization to provide theoretical guarantees for the performance of the proposed filer. Finally, motivated by a cyber-physical attack emanating from the vulnerabilities introduced by the interaction between IT infrastructure and power system, we deploy the developed theoretical results to detect such an intrusion before the functionality of the power system is disrupted.

1. INTRODUCTION

The task of FDI in control systems involves generating a diagnostic signal sensitive to the occurrence of specific faults. This task is typically accomplished by designing a filter with all available information as inputs (e.g., control signals and given measurements) and a scalar output that implements a non-zero mapping from the fault to the diagnostic signal, which is known as the residual, while decoupling unknown disturbances. The concept of residual plays a central role for the FDI problem which has been extensively studied in the last two decades.

In the context of linear systems, Beard and Jones [Bea71, Jon73] pioneered an observer-based approach whose intrinsic limitation was later improved by Massoumnia et al. [MVW89]. Following the same principles but from a game theoretic perspective, Speyer and coauthors thoroughly investigated the approach in the presence of noisy measurements [CS98, DS99]. Nyberg and Frisk extended the class of systems to linear differential-algebraic equation (DAE) apparently subsuming all the previous linear classes [NF06], which recently also studied in the context of stochastic linear systems [EFK13]. This extension greatly enhanced the applicability of FDI methods since the DAE models appear in a wide range of applications, including electrical systems, robotic manipulators, and mechanical systems.

For nonlinear systems, a natural approach is to linearize the model at an operating point, treat the nonlinear higher order terms as disturbances, and decouple their contributions from the residual by employing robust techniques [SF91, HP96]. This strategy only works well if either the system remains close to the chosen operating point, or the exact decoupling is possible. The

former approach is often limited, since in the presence of unknown inputs the system may have a wide dynamic operating range, which in case linearization leads to a large mismatch between linear model and nonlinear behavior. The latter approach was explored in detail by De Persis and Isidori, who in [PI01] proposed a differential geometric approach to extend the unobservibility subspaces of [Mas86, Section IV], and by Chen and Patton, who in [CP82, Section 9.2] dealt with a particular class of bilinear systems. These methods are, however, practically limited by the need to verify the required conditions on the system dynamics and transfer them into a standard form, which essentially involve solving partial differential equations, restricting the application of the method to relatively low dimensional systems.

Motivated by this shortcoming, in this article we develop a novel approach to FDI which strikes a balance between analytical and computational tractability, and is applicable to high dimensional nonlinear dynamics. For this purpose, we propose a design perspective that basically shifts the emphasis from the system dynamics to the family of disturbances that the system may encounter. We assume that some statistical information of the disturbance patterns is available. Following [NF06] we restrict the FDI filters to a class of linear operators that fully decouple the contribution of the linear part of the dynamics. Thanks to the linearity of the resulting filter, we then trace the contribution of the nonlinear term to the residual, and propose an optimization-based methodology to robustify the filter to the nonlinearity signatures of the dynamics by exploiting the statistical properties of the disturbance signals. The optimization formulation is effectively convex and hence tractable for high dimensional dynamics. Some preliminary results in this direction were reported in [MVAL12], while an application of our approach in the presence of measurement noise was successfully tested for wind turbines in [SMEKL13].

The performance of the proposed methodology is illustrated in an application to an emerging problem of cyber security in power networks. In modern power systems, the cyber-physical interaction of IT infrastructure (SCADA systems) with physical power systems renders the system vulnerable not only to operational errors but also to malicious external intrusions. As an example of this type of cyber-physical interaction we consider here the Automatic Generation Control (AGC) system, which is one of the few control loops in power networks that are closed over the SCADA system without human operator intervention. In earlier work [MVM+10, MVM+11] we have shown that, having gained access to the AGC signal, an attacker can provoke frequency deviations and power oscillations by applying sophisticated attack signals. The resulting disruption can be serious enough to trigger generator out-of-step protection relays, leading to load shedding and generator tripping. Our earlier work, however, also indicated that an early detection of the intrusion may allow one to disconnect the AGC and limit the damage by relying solely on the so-called primary frequency controllers. In this work we show how to mitigate this cyber-physical security concern by using the proposed FDI scheme to develop a protection layer which quickly detects the abnormal signals generated by the attacker. This approach to enhancing the cyber-security of power transmission systems led to an EU patent sponsored by ETH Zurich [MEVAL].

The article is organized as follows. In Section 2 a formal description of the FDI problem as well as the outline of the proposed methodology is presented. A general class of nonlinear models is described in Section 3. Then, reviewing residual generation for the linear models, we develop an optimization-based framework for nonlinear systems in Section 4. Theoretical guarantees are also provided in the context of randomized algorithms. We apply the developed methodology to the AGC case study in Section 5, and finally conclude with some remarks and directions for
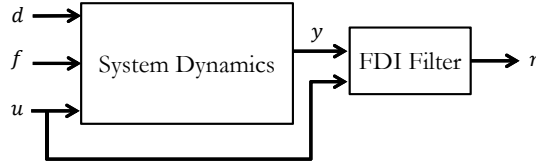
FIGURE 1. General configuration of the FDI filter

future work in Section 6. For better readability, the technical proofs of Sections 4.2 and 4.3 are moved to the appendices.

Notation. The symbols $\mathbb{N}$ and $\mathbb{R}_+$ denote the set of natural and nonnegative real numbers, respectively. Let $A \in \mathbb{R}^{n \times m}$ be an $n \times m$ matrix with real values, $A^{\mathsf{T}} \in \mathbb{R}^{m \times n}$ be its transpose, and $\|A\|_2 := \overline{\sigma}(A)$ where $\overline{\sigma}$ is the maximum singular value of the matrix. Given a vector $v := [v_1, \cdots, v_n]^{\mathsf{T}}$, the infinite norm is defined as $\|v\|_\infty := \max_{i \le n} |v_i|$. Let G be a linear matrix transfer function. Then $\|G\|_{\mathcal{H}_\infty} := \sup_{\omega \in \mathbb{R}} \overline{\sigma}\big(G(j\omega)\big)$, where $\overline{\sigma}$ is the maximum singular value of the matrix $G(j\omega)$. The function space $\mathcal{W}^n$ denotes the set of piece-wise continuous (p.w.c) functions taking values in $\mathbb{R}^n$, and $\mathcal{W}_T^n$ is the restriction of $\mathcal{W}^n$ to the time interval $[0, T]$, which is endowed with the $\mathcal{L}_2$-inner product, i.e., $\langle e_1, e_2 \rangle := \int_0^T e_1^{\mathsf{T}}(t) e_2(t) \mathrm{d}t$ with the associated $\mathcal{L}_2$-norm $\|e\|_{\mathcal{L}_2} := \sqrt{\langle e, e \rangle}$. The linear operator $p : \mathcal{W}^n \to \mathcal{W}^n$ is the distributional derivative operator. In particular, if $e : \mathbb{R}_+ \to \mathbb{R}^n$ is a smooth mapping then $p[e(t)] := \frac{\mathrm{d}}{\mathrm{d}t} e(t)$. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we denote the n-Cartesian product space by $\Omega^n := \bigotimes_{i=1}^n \Omega$ and the respective product measure by $\mathbb{P}^n$.

2. PROBLEM STATEMENT AND OUTLINE OF THE PROPOSED APPROACH

In this section, we provide the formal description of the FDI problem as well as our new design perspective. We will also outline our methodology to tackle the proposed perspective.

2.1. **Formal Description.** The objective of the FDI design is to use all information to generate a diagnostic signal to alert the operators to the occurrence of a specific fault. Consider a general dynamical system as in Figure 1 with its inputs categorized into (i) unknown inputs d, (ii) fault signal f, and (iii) known inputs u. The unknown input d represents unknown disturbances that the dynamical system encounters during normal operation. The known input u contains all known signals injected to the system which together with the measurements y are available for FDI tasks. Finally, the input f is a fault (or an intrusion) which cannot be directly measured and represents the signal to be detected.

The FDI task is to design a filter whose input are the known signals (u and y) and whose output (known as the residual and denoted by r) differentiates whether the measurements are a consequence of some normal disturbance input d, or due to the fault signal f. Formally speaking, the residual can be viewed as a function $r(d, f)$, and the FDI design is ideally translated as the mapping requirements

$$d \mapsto r(d, 0) \equiv 0, \tag{1a}$$

$$f \mapsto r(d, f) \neq 0, \quad \forall d \tag{1b}$$

where condition (1a) ensures that the residual of the filter, r, is not excited when the system is perturbed by normal disturbances d, while condition (1b) guarantees the filter sensitivity to the fault f in the presence of any disturbance d.

The state of the art in FDI concentrates on the system dynamics, and imposes restrictions to provide theoretical guarantees for the required mapping conditions (1). For example, the authors in [NF06] restrict the system to linear dynamics, whereas [HKEY99, PI01] treat nonlinear systems but impose necessary conditions in terms of a certain distribution connected to their dynamics. In an attempt to relax the perfect decoupling condition, one may consider the worst case scenario of the mapping (1) in a robust formulation as

$$
(2) \qquad \mathrm{RP} : \begin{cases} \min_{\gamma, \mathfrak{F}} & \gamma \\ \text{s.t.} & \big\| r(d, 0) \big\| \leq \gamma, \qquad \forall d \in \mathcal{D} \\ & f \mapsto r(d, f) \neq 0, \quad \forall d \in \mathcal{D}, \end{cases}
$$

where $\mathcal{D}$ is set of normal disturbances, γ is the alarm threshold of the designed filter, and the minimization is running over a given class of FDI filters denoted by $\mathfrak{F}$. Note that the residual r is influenced by the choice of the filter in $\mathfrak{F}$, but we omit this dependence for notational simplicity. In view of formulation (2), an alarm is only raised whenever the residual exceeds γ, i.e., the filter avoids any false alarm. This, however, comes at the cost of missed detections of the faults whose residual is not bigger than the threshold γ. In the literature, the robust perspective RP has also been studied in order for a trade-off between disturbance rejection and fault sensitivity for a certain class of dynamics, e.g., see [CP82, Section 9.2] for bilinear dynamics and [FF12] for multivariate polynomial systems.

2.2. **New Design Perspective.** Here we shift our attention from the system dynamics to the class of unknown inputs $\mathcal{D}$. We assume that the disturbance signal d comes from a prescribed probability space and relax the robust formulation RP by introducing probabilistic constraints instead. In this view, the performance of the FDI filter is characterized in a probabilistic fashion.

Assume that the signal d is modeled as a random variable on the prescribed probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which takes values in a metric space endowed with the corresponding Borel sigma-algebra. Assume further that the class of FDI filters ensures the measurability of the mapping $d \mapsto r$ where r also belongs to a metric space. In light of this probabilistic framework, one may quantify the filter performance from different perspectives; in the following we propose two of them:

$$
(3) \quad \mathrm{AP} : \begin{cases} \min_{\gamma, \mathfrak{F}} & \gamma \\ \text{s.t.} & \mathbb{E}\big[J\big(\| r(d, 0) \| \big) \big] \leq \gamma \\ & f \mapsto r(d, f) \neq 0, \quad \forall d \in \mathcal{D}, \end{cases} \qquad \mathrm{CP} : \begin{cases} \min_{\gamma, \mathfrak{F}} & \gamma \\ \text{s.t.} & \mathbb{P}\big(\| r(d, 0) \| \leq \gamma \big) \geq 1 - \varepsilon \\ & f \mapsto r(d, f) \neq 0, \quad \forall d \in \mathcal{D}, \end{cases}
$$

where $\mathbb{E}[\,\cdot\,]$ in AP is meant with respect to the probability measure $\mathbb{P}$, and $\| \cdot \|$ is the corresponding norm in the r space. The function $J : \mathbb{R}_+ \to \mathbb{R}_+$ in AP and $\varepsilon \in (0, 1)$ in CP are design parameters. To control the filter residual generated by d, the payoff function J is required to be in class $\mathcal{K}_\infty$, i.e., J is strictly increasing and $J(0) = 0$ [Kha92, Definition 4.2, p. 144]. The decision variables in the above optimization programs are $\mathfrak{F}$, a class of FDI filters which is chosen a priori, and γ which is the filter threshold; we shall explain these design parameters more explicitly in subsequent sections.

Two formulations provide different probabilistic interpretations of fault detection. The program AP stands for "*Average Performance*" and takes all possible disturbances into account, but in accordance with their occurrence probability in an averaging sense. The program CP stands for "*Chance Performance*" and ignores an ε-fraction of the disturbance patterns and only aims to optimize the performance over the rest of the disturbance space. Note that in the CP perspective, the parameter ε is an additional design parameter to be chosen a priori.

Let us highlight that the proposed perspectives rely on the probability distribution $\mathbb{P}$, which requires prior information about possible disturbance patterns. That is, unlike the existing literature, the proposed design prioritizes between disturbance patterns in terms of their occurrence likelihood. From a practical point of view this requirement may be natural; in Section 5 we will describe an application of this nature.

2.3. **Outline of the Proposed Methodology.** We employ randomized algorithms to tackle the formulations in (3). We generate n independent and identically distributed (i.i.d.) scenarios $(d_i)_{i=1}^n$ from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and consider the following optimization problems as random counterparts of those in (3):

$$(4) \quad \widetilde{\mathrm{AP}} : \begin{cases} \min_{\gamma, \mathfrak{F}} & \gamma \\ \text{s.t.} & \frac{1}{n} \sum_{i=1}^n J\big(\|r(d_i, 0)\|\big) \leq \gamma \\ & f \mapsto r(d, f) \neq 0, \quad \forall d \in \mathcal{D} \end{cases} \qquad \widetilde{\mathrm{CP}} : \begin{cases} \min_{\gamma, \mathfrak{F}} & \gamma \\ \text{s.t.} & \max_{i \leq n} \|r(d_i, 0)\| \leq \gamma \\ & f \mapsto r(d, f) \neq 0, \quad \forall d \in \mathcal{D}, \end{cases}$$

Notice that the optimization problems $\widetilde{\mathrm{AP}}$ and $\widetilde{\mathrm{CP}}$ are naturally stochastic as they depend on the generated scenarios $(d_i)_{i=1}^n$, which is indeed a random variable defined on n-fold product probability space $(\Omega^n, \mathcal{F}^n, \mathbb{P}^n)$. Therefore, their solutions are also random variables. In this work, we first restrict the FDI filters to a class of linear operators in which the random programs (4) are effectively convex, and hence tractable. In this step, the FDI filter is essentially robustified to n signatures of the dynamic nonlinearity. Subsequently, invoking existing results on randomized optimization, in particular [Han12, MSL15], we will provide probabilistic guarantees on the relation of programs (3) and their probabilistic counterparts in (4), whose precision is characterized in terms of the number of scenarios n.

We should highlight that the true empirical approximation of the chance constraint in CP is indeed $\frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\left\{ \|r(d_i, 0)\| \leq \gamma \right\}} \geq 1 - \varepsilon$, where $\mathbb{1}$ is the indicator function. This approximation, as opposed to the one proposed in (4), leads to a non-convex optimization program which is, in general, computationally intractable. In addition, note that the design parameter ε of CP in (3) does not explicitly appear in the random counterpart $\widetilde{\mathrm{CP}}$ in (4). However, as we will clarify in 4.3, the parameter ε contributes to the probabilistic guarantees of the design.

3. Model Description and Basic Definitions

In this section we introduce a class of nonlinear models along with some basic definitions, which will be considered as the system dynamics in Figure 1 throughout the article. Consider the nonlinear differential-algebraic equation (DAE) model

$$(5) \qquad E(x) + H(p)x + L(p)z + F(p)f = 0,$$

where the signals x, z, f are assumed to be piece-wise continuous (p.w.c.) functions from $\mathbb{R}_+$ into $\mathbb{R}^{n_x}, \mathbb{R}^{n_z}, \mathbb{R}^{n_f}$, respectively; we denote the spaces of such signals by $\mathcal{W}^{n_x}, \mathcal{W}^{n_z}, \mathcal{W}^{n_f}$, respectively.

Let n_r be the number of rows in (5), and $E : \mathbb{R}^{n_x} \to \mathbb{R}^{n_r}$ be a Lipschitz continuous mapping. The operator p is the distributional derivative operator [Ada75, Section I], and H, L, F are polynomial matrices in the operator p with n_r rows and n_x, n_z, n_f columns, respectively. In the setup of Figure 1, the signal x represents all unknowns signals, e.g., internal states of the system dynamics and unknown disturbances d. The signal z contains all known signals, i.e., it is an augmented signal including control input u and available measurements y. The signal f stands for faults or intrusion which is the target of detection. We refer to [Shc07] and the references therein for general theory of nonlinear DAE systems and the regularity of their solutions.

One may extend the space of functions x, z, f to Sobolev spaces, but this is outside the scope of our study. On the other hand, if these spaces are restricted to the (resp. right) smooth functions, then the operator p can be understood as the classical (resp. right) differentiation operator. Throughout this article we will focus on continuous-time models, but one can obtain similar results for discrete-time models by changing the operator p to the time-shift operator. We will think of the matrices $H(p)$, $L(p)$ and $F(p)$ above either as linear operators on the function spaces (in which case p will be interpreted as a generalized derivative operator as explained above) or as algebraic objects (in which case p will be interpreted as simply a complex variable). The reader is asked to excuse this slight abuse of the notation, but the interpretation should be clear from the context.

Let us first show the generality of the DAE framework of (5) by the following example. Consider the classical nonlinear ordinary differential equation

(6)
$$\begin{cases} G\dot{X}(t) &= E_X\big(X(t), d(t)\big) + AX(t) + B_u u(t) + B_d d(t) + B_f f(t) \\ Y(t) &= E_Y\big(X(t), d(t)\big) + CX(t) + D_u u(t) + D_d d(t) + D_f f(t) \end{cases}$$

where $u(\cdot)$ is the input signal, $d(\cdot)$ the unknown disturbance, $Y(\cdot)$ the measured output, $X(\cdot)$ the internal variables, and $f(\cdot)$ a faults (or an attack) signal to be detected. Parameters G, A, B_u, B_d, B_f, D_u, D_d, and D_f are constant matrices and functions E_X, E_Y are Lipschitz continuous mappings with appropriate dimensions. One can easily fit the model (6) into the DAE framework of (5) by defining

$$x := \begin{bmatrix} X \\ d \end{bmatrix}, \qquad z := \begin{bmatrix} Y \\ u \end{bmatrix},$$

$$E(x) := \begin{bmatrix} E_X(x) \\ E_Y(x) \end{bmatrix}, \quad H(p) := \begin{bmatrix} -pG + A & B_d \\ C & D_d \end{bmatrix}, \quad L(p) := \begin{bmatrix} 0 & B_u \\ -I & D_u \end{bmatrix}, \quad F(p) := \begin{bmatrix} B_f \\ D_f \end{bmatrix}.$$

Following [NF06], with a slight extension to a nonlinear dynamics, let us formally characterize all possible observations of the model (5) in the absence of the fault signal f:

(7)
$$\mathcal{M} := \big\{ z \in \mathcal{W}^{n_z} \,\big|\, \exists x \in \mathcal{W}^{n_x} : \quad E(x) + H(p)x + L(p)z = 0 \big\};$$

This set is known as the *behavior* of the system [PW98].

Definition 3.1 (Residual Generator)**.** *A proper linear time invariant filter $r := R(p)z$ is a residual generator for* (5) *if for all $z \in \mathcal{M}$, it holds that $\lim_{t \to \infty} r(t) = 0$.*

Note that by Definition 3.1 the class of residual generators in this study is restricted to a class of *linear* transfer functions where $R(p)$ is a matrix of proper rational functions of p.

Definition 3.2 (Fault Sensitivity). *The residual generator introduced in Definition 3.1 is sensitive to fault f_i if the transfer function from f_i to r is nonzero, where f_i is the ith elements of the signal f.*

One can inspect that Definition 3.1 and Definition 3.2 essentially encode the basic mapping requirements (1a) and (1b), respectively.

4. Fault Detection and Isolation Filters

The main objective of this section is to establish a scalable framework geared towards the design perspectives AP and CP as explained in Section 2. To this end, we first review a polynomial characterization of the residual generators and its linear program formulation counterpart for linear systems (i.e., the case where $E(x) \equiv 0$). We then extend the approach to the nonlinear model (5) to account for the contribution of $E(\cdot)$ to the residual, and subsequently provide probabilistic performance guarantees for the resulting filter.

4.1. **Residual Generators for Linear Systems.** In this subsection we assume $E(x) \equiv 0$, i.e., we restrict our attention to the class of linear DAEs. One can observe that the behavior set $\mathcal{M}$ can alternatively be defined as

$$\mathcal{M} = \big\{ z \in \mathcal{W}^{n_z} \,\big|\, N_H(p)L(p)z = 0 \big\},$$

where the collection of the rows of $N_H(p)$ forms an irreducible polynomial basis for the left null-space of the matrix $H(p)$ [PW98, Section 2.5.2]. This representation allows one to describe the residual generators in terms of polynomial matrix equations. That is, by picking a linear combination of the rows of $N_H(p)$ and considering an arbitrary polynomial $a(p)$ of sufficiently high order with roots with negative real parts, we arrive at a residual generator in the sense of Definition 3.1 with transfer operator

$$(8) \qquad R(p) = a^{-1}(p)\gamma(p)N_H(p)L(p) := a^{-1}(p)N(p)L(p),$$

where $\gamma(p)$ is a polynomial row vector representing a linear combination of the rows of $N_H(p)$. Note that the role of $\gamma(p)$ is implicitly taken into consideration by $N(p) := \gamma(p)N_H(p)$. The above filter can easily be realized by an explicit state-space description with input z and output r. Multiplying the left hand-side of (5) by $a^{-1}(p)N(p)$ leads to

$$r = -a^{-1}(p)N(p)F(p)f.$$

Thus, a sensitive residual generator, in the sense of Definition 3.1 and Definition 3.2, is characterized by the polynomial matrix equations

$$(9a) \qquad\qquad\qquad\qquad N(p)H(p) = 0,$$
$$(9b) \qquad\qquad\qquad\qquad N(p)F(p) \neq 0,$$

where (9a) implements condition (1a) above (cf. Definition 3.1) while (9b) implements condition (1b) (cf. Definition 3.2). Both row polynomial vector $N(p)$ and denominator polynomial $a(p)$ can be viewed as design parameters. Throughout this study we, however, fix $a(p)$ and aim to find an optimal $N(p)$ with respect to a certain objective criterion related to the filter performance.

In case there are more than one faults ($n_f > 1$), it might be of interest to isolate the impact of one fault in the residual from the others. The following remark implies that the isolation problem is effectively a detection problem.

Remark 4.1 (Fault Isolation)**.** *Consider model* (5) *and suppose* $n_f > 1$. *In order to detect only one of the fault signals, say* f_1, *and isolate it from the other faults,* $f_i, i \in \{2, \cdots, n_f\}$, *one may consider the detection problem for the same model but in new representation*

$$E(x) + [H(p) \ \widetilde{F}(p)] \begin{bmatrix} x \\ \tilde{f} \end{bmatrix} + L(p)z + F_1(p)f = 0,$$

where $F_1(p)$ *is the first column of* $F(p)$, *and* $\widetilde{F}(p) := [F_2(p), \cdots, F_{n_f}(p)]$, *and* $\tilde{f} := [f_2, \cdots, f_{n_f}]$.

In light of Remark 4.1, one can build a bank of filters where each filter aims to detect a particular fault while isolating the impact of the others; see [FKA09, Theorem 2] for more details on fault isolation. Next, we show how to transform the matrix polynomial equations (9) into a linear programming framework.

Lemma 4.2. *Let* $N(p)$ *be a feasible polynomial matrix of degree* d_N *for the inequalities* (9), *where*

$$H(p) := \sum_{i=0}^{d_H} H_i p^i, \quad F(p) := \sum_{i=0}^{d_F} F_i p^i, \quad N(p) := \sum_{i=0}^{d_N} N_i p^i,$$

and $H_i \in \mathbb{R}^{n_r \times n_x}$, $F_i \in \mathbb{R}^{n_r \times n_f}$, *and* $N_i \in \mathbb{R}^{1 \times n_r}$ *are constant matrices. Then, the polynomial matrix inequalities* (9) *are equivalent, up to a scalar, to*

(10a) $$\bar{N}\bar{H} = 0,$$

(10b) $$\left\| \bar{N}\bar{F} \right\|_\infty \geq 1,$$

where $\| \cdot \|_\infty$ *is the infinity vector norm, and*

$$\bar{N} := \begin{bmatrix} N_0 & N_1 & \cdots & N_{d_N} \end{bmatrix}$$

$$\bar{H} := \begin{bmatrix} H_0 & H_1 & \cdots & H_{d_H} & 0 & \cdots & 0 \\ 0 & H_0 & H_1 & \cdots & H_{d_H} & 0 & \vdots \\ \vdots & & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & 0 & H_0 & H_1 & \cdots & H_{d_H} \end{bmatrix},$$

$$\bar{F} := \begin{bmatrix} F_0 & F_1 & \cdots & F_{d_F} & 0 & \cdots & 0 \\ 0 & F_0 & F_1 & \cdots & F_{d_F} & 0 & \vdots \\ \vdots & & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & 0 & F_0 & F_1 & \cdots & F_{d_F} \end{bmatrix}.$$

Proof. It is easy to observe that

$$N(p)H(p) = \bar{N}\bar{H}[I \ pI \ \cdots \ p^i I]^\mathsf{T}, \qquad i := d_N + d_H,$$
$$N(p)F(p) = \bar{N}\bar{F}[I \ pI \ \cdots \ p^j I]^\mathsf{T}, \qquad j := d_N + d_F.$$

Moreover, in light of the linear structure of equations (9), one can simply scale the inequality (9b) and arrive at the assertion of the lemma. □

Strictly speaking, the formulation in Lemma 4.2 is not a linear program, due to the non-convex constraint (10b). It is, however, easy to show that the characterization (10) can be understood as a number of linear programs, which grows linearly in the degree of the filter:

Lemma 4.3. *Consider the sets*

$$\mathcal{N}_j := \left\{ \bar{N} \in \mathbb{R}^{n_r(d_N+1)} \mid \bar{N}\bar{H} = 0, \ \bar{N}\bar{F}v_j \geq 1 \right\}, \qquad v_j := \left[0, \cdots, \overset{\downarrow \ j^{th}}{1}, \cdots, 0\right]^\mathsf{T},$$

and let $\mathcal{N} := \bigcup_{j=1}^m \mathcal{N}_j$ where $m := n_f(d_F + d_N + 1)$ is the number of columns of $\bar{F}$ (the parameters $\bar{H}, \bar{F}, n_f, d_F, d_N$ are as considered in Lemma 4.2). Then, the set characterized by (10) is equivalent to $\mathcal{N} \cup -\mathcal{N}$.

Proof. Notice that $\|\bar{N}\bar{F}\|_\infty \geq 1$ if and only if there exists a coordinate j such that $\bar{N}\bar{F}v_j \geq 1$ or $\bar{N}\bar{F}v_j \leq -1$. Thus, the proof readily follows from the fact that each of the set $\mathcal{N}_j$ focuses on a component of the vector $\bar{N}\bar{F}$ in (10b). $\qquad\square$

Fact 4.4. *There exists a solution $N(p)$ to (9) if and only if Rank $[H(p) \ F(p)] > $ Rank $H(p)$.*

Fact 4.4 provides necessary and sufficient conditions for the feasibility of the linear program formulation in Lemma 4.2; proof is omitted as it is an easy adaptation of the one in [FKA09, Corollary 3].

4.2. **Extension to Nonlinear Systems.** In the presence of nonlinear terms $E(x) \neq 0$, it is straightforward to observe that the residual of filter (8) consists of two terms:

$$(11) \qquad r := R(p)z = -\underbrace{a^{-1}(p)N(p)F(p)f}_{(i)} - \underbrace{a^{-1}(p)N(p)E(x)}_{(ii)}.$$

Term (i) is the desired contribution of the fault f and is in common with the linear setup. Term (ii) is due to the nonlinear term $E(\cdot)$ in (5). Our aim here will be to reduce the impact of $E(x)$ while increasing the sensitivity to the fault f. To achieve this objective, we develop two approaches to control each of the two terms separately; in both cases we assume that the degree of the filter (i.e., d_N in Lemma 4.2) and the denominator (i.e., $a(p)$ in (11)) are fixed, and the aim is to design the numerator coefficients (i.e., $N(p)$ in (11)).

Approach (I) (Fault Sensitivity). To focus on fault sensitivity while neglecting the contribution of the nonlinear term, we assume that the system operates close to an equilibrium point $x_e \in \mathbb{R}^{n_x}$. Even though in case of a fault the system may eventually deviate substantially from its nominal operating point, if the FDI filter succeeds in identifying the fault early the system will not have time to deviate too far. Hence, one may hope that a filter based on linearizing the system dynamics around the equilibrium would suffice. Then we assume, without loss of generality, that

$$\lim_{x \to x_e} \frac{\|E(x)\|_2}{\|x - x_e\|_2} = 0,$$

where $\|\cdot\|_2$ stands for the Euclidean norm of a vector. If this is not the case, the linear part of $E(\cdot)$ can be extracted and included in the linear part of the system.

To increase the sensitivity of the linear filter to the fault f, we revisit the linear programming formulation (10) and seek a feasible numerator $N(p)$ such that the coefficients of the transfer function $N(p)F(p)$ attain maximum values within the admissible range. This gives rise to the

following optimization problem:

$$
(12) \qquad
\begin{cases}
\max\limits_{\bar{N}} & \|\bar{N}\bar{F}\|_{\infty} \\
\text{s.t.} & \bar{N}\bar{H} = 0 \\
& \|\bar{N}\|_{\infty} \leq 1
\end{cases}
$$

where the objective function targets the contribution of the signal f to the residual r. Let us recall that $\bar{N}\bar{F}$ is the vector containing all numerator coefficients of the transfer function $f \mapsto r$. The second constraint in (12) is added to ensure that the solutions remain bounded; note that thanks to the linearity of the filter this constraint does not influence the performance. Though strictly speaking (12) is not a linear program, in a similar fashion as in Lemma 4.3 it is easy to transform it to a family of m different linear programs, where m is the number of columns of $\bar{F}$.

How well the filter designed by (12) will work depends on the magnitude of the second term in (11), which is due to the nonlinearities $E(x)$ and is ignored in (12). If the term generated by $E(x)$ is large enough, the filter may lead to false alarms, whereas if we set our thresholds high to tolerate the disturbance generated by $E(x)$ in nominal conditions, the filter may lead to missed detections. A direct way toward controlling this trade-off involving the nonlinear term will be the focus of the second approach.

Approach (II) (Robustify to Nonlinearity Signatures). This approach is the main step toward the theoretical contribution of the article, and provides the principle ingredients to tackle the proposed perspectives AP and CP introduced in (3). The focus is on term (ii) of the residual (11), in relation to the mapping (1a). The idea is to robustify the filter against certain signatures of the nonlinearity during nominal operation. In the following we restrict the class of filters to the feasible solutions of polynomial matrix equations (9), characterized in Lemma 4.2.

Let us denote the space of all p.w.c. functions from the interval $[0, T]$ to $\mathbb{R}^n$ by $\mathcal{W}_T^n$. We equip this space with the $\mathcal{L}_2$-inner product and the corresponding norm

$$
\|e\|_{\mathcal{L}_2} := \sqrt{\langle e, e \rangle}, \qquad \langle e, g \rangle := \int_0^T e^{\mathsf{T}}(t) g(t) \mathrm{d}t, \quad e, g \in \mathcal{W}_T^n.
$$

Consider an unknown signal $x \in \mathcal{W}_T^{n_x}$. In the context of the ODEs (6) that means we excite the system with the disturbance $d(\cdot)$ for the time horizon T. We then stack $d(\cdot)$ together with the internal state $X(\cdot)$ to introduce $x := \begin{bmatrix} X \\ d \end{bmatrix}$. We define the signals $e_x \in \mathcal{W}_T^{n_r}$ and $r_x \in \mathcal{W}_T^1$ as follows:

$$
(13) \qquad e_x(t) := E\big(x(t)\big), \qquad r_x(t) := -a^{-1}(p)N(p)[e_x](t), \qquad \forall t \in [0, T].
$$

The signal e_x is the "*nonlinearity signature*" in the presence of the unknown signal x, and the signal r_x is the contribution of the nonlinear term to the residual of the linear filter. Our goal now is to minimize $\|r_x\|_{\mathcal{L}_2}$ in an optimization framework in which the coefficients of polynomial $N(p)$ are the decision variables and the denominator $a(p)$ is a fixed stable polynomial with the degree at least the same as $N(p)$.

Lemma 4.5. *Let $N(p)$ be a polynomial row vector of dimension n_r and degree d_N, and $a(p)$ be a stable scalar polynomial with the degree at least d_N. For any $x \in \mathcal{W}_T^{n_x}$ there exists $\psi_x \in \mathcal{W}_T^{n_r(d_N+1)}$ such that*

$$
(14a) \qquad r_x(t) = \bar{N}\psi_x(t), \qquad \forall t \in [0, T]
$$

$$
(14b) \qquad \|\psi_x\|_{\mathcal{L}_2} \leq C \|e_x\|_{\mathcal{L}_2}, \qquad C := \sqrt{n_r(d_N+1)}\|a^{-1}\|_{\mathcal{H}_{\infty}},
$$

where $\bar{N}$ is the vector collecting all the coefficients of the numerator $N(p)$ as introduced in Lemma 4.2, and the signals e_x and r_x are defined as in (13).

Proof. See Appendix I.1. $\qquad\square$

Given $x \in \mathcal{W}_T^{n_x}$ and the corresponding function ψ_x as defined in Lemma 4.5, we have

$$(15) \qquad \|r_x\|_{\mathcal{L}_2}^2 = \bar{N} Q_x \bar{N}^\intercal, \qquad Q_x := \int_0^T \psi_x(t)\psi_x^\intercal(t)\mathrm{d}t.$$

We call Q_x the "*signature matrix*" of the nonlinearity signature $t \mapsto e_x(t)$ resulting from the unknown signal x. Given x and the corresponding signature matrix Q_x, the $\mathcal{L}_2$-norm of r_x in (13) can be minimized by considering an objective which is a quadratic function of the filter coefficients $\bar{N}$ subject to the linear constraints in (10):

$$(16) \qquad \begin{cases} \min\limits_{\bar{N}} & \bar{N} Q_x \bar{N}^\intercal \\ \text{s.t.} & \bar{N}\bar{H} \quad = 0 \\ & \left\|\bar{N}\bar{F}\right\|_\infty \geq 1 \end{cases}$$

The program (16) is not a true quadratic program due to the second constraint. Following Lemma 4.3, however, one can show that the optimization program (16) can be viewed as a family of m quadratic programs where $m = n_f(d_F + d_N + 1)$.

In the rest of the subsection, we establish an algorithmic approach to approximate the matrix Q_x for a given $x \in \mathcal{W}_T^{n_x}$, with an arbitrary high precision. We first introduce a finite dimensional subspace of $\mathcal{W}_T^1$ denoted by

$$(17) \qquad \mathcal{B} := span\{b_0, b_1, \cdots, b_k\},$$

where the collection of $b_i : [0,T] \to \mathbb{R}$ is a basis for $\mathcal{B}$. Let $\mathcal{B}^{n_r} := \bigotimes_{i=1}^{n_r} \mathcal{B}$ be the n_r Cartesian product of the set $\mathcal{B}$, and $\mathbb{T}_\mathcal{B} : \mathcal{W}_T^{n_r} \to \mathcal{B}^{n_r}$ be the $\mathcal{L}_2$-orthogonal projection operator onto $\mathcal{B}^{n_r}$, i.e.,

$$(18) \qquad \mathbb{T}_\mathcal{B}(e_x) = \sum_{i=0}^k \beta_i^\star b_i, \qquad \beta^\star := \arg\min_\beta \left\|e_x - \sum_{i=0}^k \beta_i b_i\right\|_{\mathcal{L}_2}$$

Let us remark that if the basis of $\mathcal{B}$ is orthonormal (i.e., $\langle b_i, b_j \rangle = 0$ for $i \neq j$), then $\beta_i^\star = \int_0^T b_i(t)e_x(t)\mathrm{d}t$; we refer to [Lue69, Section 3.6] for more details on the projection operator.

Assumption 4.6. *We stipulate that*

(i) *The basis functions b_i of subspace $\mathcal{B}$ are smooth and $\mathcal{B}$ is closed under the differentiation operator p, i.e., for any $b \in \mathcal{B}$ we have $p[b] = \frac{\mathrm{d}}{\mathrm{d}t} b \in \mathcal{B}$.*

(ii) *The basis vectors in* (17) *are selected from an $\mathcal{L}_2$-complete basis for $\mathcal{W}_T^1$, i.e., for any $e \in \mathcal{W}_T^{n_r}$, the projection error $\left\|e - \mathbb{T}_\mathcal{B}(e)\right\|_{\mathcal{L}_2}$ can be made arbitrarily small by increasing the dimension k of subspace $\mathcal{B}$.*

The requirements of Assumptions 4.6 can be fulfilled for subspaces generated by, for example, the polynomial or Fourier basis. Thanks to Assumption 4.6(i), the linear operator p can be

viewed as a matrix operator. That is, there exists a square matrix D with dimension $k+1$ such that

$$(19) \qquad p[B(t)] = \frac{\mathrm{d}}{\mathrm{d}t} B(t) = D B(t), \qquad B(t) := [b_0(t), \cdots, b_k(t)]^{\mathsf{T}}.$$

In Section 5.2 we will provide an example of such matrix operator for the Fourier basis. By virtue of the matrix representations of (19) we have

$$(20) \qquad N(p) \mathbb{T}_{\mathcal{B}}(e_x) = \sum_{i=0}^{d_N} N_i p^i \beta^\star B = \sum_{i=0}^{d_N} N_i \beta^\star D^i B = \bar{N} \bar{D} B, \qquad \bar{D} := \begin{bmatrix} \beta^\star \\ \beta^\star D \\ \vdots \\ \beta^\star D^{d_N} \end{bmatrix},$$

where the vector $\beta^\star := [\beta_0^\star, \cdots, \beta_k^\star]$ is introduced in (18). If we define the positive semidefinite matrix $G := [G_{ij}]$ of dimension $k+1$ by

$$(21) \qquad G_{ij} := \left\langle a^{-1}(p)[b_i], a^{-1}(p)[b_j] \right\rangle,$$

we arrive at

$$(22) \qquad \left\| a^{-1}(p) N(p) \mathbb{T}_{\mathcal{B}}(e) \right\|_{\mathcal{L}_2}^2 = \bar{N} Q_{\mathcal{B}} \bar{N}^{\mathsf{T}}, \qquad Q_{\mathcal{B}} := \bar{D} G \bar{D}^{\mathsf{T}},$$

where $\bar{D}$ and G are defined in (20) and (21), respectively. Note that the matrices G and D are built by the data of the subspace $\mathcal{B}$ and denominator $a(p)$, whereas the nonlinearity signature only influences the coefficient $\beta^\star$. The above discussion is summarized in Algorithm 1 with an emphasis on models described by the ODE (6), while Proposition 4.7 addresses the precision of the approximation scheme.

Proposition 4.7 (Signature Matrix Approximation). *Consider an unknown signal $x : [0, T] \to \mathbb{R}^{n_x}$ in $\mathcal{W}_T^{n_x}$ and the corresponding nonlinearity signature e_x and signature matrix Q_x as defined in (13) and (15), respectively. Let $(b_i)_{i \in \mathbb{N}} \subset \mathcal{W}_T^1$ be a family of basis functions satisfying Assumptions 4.6, and let $\mathcal{B}$ be the finite dimensional subspace in (17). If $\|e_x - \mathbb{T}_{\mathcal{B}}(e_x)\|_{\mathcal{L}_2} < \delta$, where $\mathbb{T}_{\mathcal{B}}$ is the projection operator onto $\mathcal{B}^{n_r}$, then*

$$(23) \qquad \left\| Q_x - Q_{\mathcal{B}} \right\|_2 < \bar{C}\delta, \qquad \bar{C} := \left(1 + 2\|e_x\|_{\mathcal{L}_2} \right) C \|a^{-1}\|_{\mathcal{H}_\infty},$$

where $Q_{\mathcal{B}}$ is obtained by (22) (the output of Algorithm 1), and C is the same constant as in (14b).

Proof. See Appendix I.1. $\qquad \square$

Remark 4.8 (Multi Signatures Training). *In order to robustify the FDI filter to more than one unknown signal, say $\{x_i(\cdot)\}_{i=1}^n$, one may introduce an objective function as an average cost $\bar{N} \left(\frac{1}{n} \sum_{i=1}^n Q_{x_i} \right) \bar{N}^{\mathsf{T}}$ or the worst case viewpoint $\max_{i \le n} \bar{N} Q_{x_i} \bar{N}^{\mathsf{T}}$, where Q_{x_i} is the signature matrix corresponding to x_i as defined in (15).*

4.3. **Proposed Methodology and Probabilistic Performance.** The preceding subsection proposed two optimization-based approaches to enhance the FDI filter design from linear to nonlinear system dynamics. Approach (I) targets the fault sensitivity while neglecting the nonlinear term of the system dynamics, and Approach (II) offers a QP framework to robustify the residual with respect to signatures of the dynamic nonlinearities. Here our aim is to achieve a

---

Algorithm 1 Computing the signature matrix Q_x in (15)

(i) **Initialization of the Filter Paramters:**

    (a) Select a stable filter denominator $a(p)$, a numerator degree d_N not higher than $a(p)$ order, and horizon T

    (b) Select a basis $\{b_i\}_{i=1}^k \subset \mathcal{W}_T^1$ satisfying Assumptions 4.6

    (c) Compute the differentiation matrix D in (19)

    (d) Compute the matrix G in (21) [1]

(ii) **Identification of the Nonlinearity Signature:**

    (a) <u>Input</u> the disturbance pattern $d(\,\cdot\,)$ for time horizon T

    (b) Solve (6) under inputs $d(\,\cdot\,)$ and $f \equiv 0$ to obtain the internal state $X(\,\cdot\,)$

    (c) Set the unknown signal $x(t) := [X^\intercal(t),\ d^\intercal(t)]^\intercal$

    (d) Set the nonlinearity signature $e_x(t) := \left[E_X^\intercal\big(x(t)\big),\ E_Y^\intercal\big(x(t)\big)\right]^\intercal$

(iii) **Computation of the Signature Matrix**

    (a) Compute $\beta^\star$ from (18) (in case of orthonormal basis $\beta_i^\star = \int_0^T b_i(t)e_x(t)\mathrm{d}t$)

    (b) Compute $\bar{D}$ from (20)

    (c) <u>Ouput</u> $Q_{\mathcal{B}} := \bar{D}G\bar{D}^\intercal$ in (22)

---

reconciliation between these two approaches. We subsequently provide theoretical results from the proposed solutions to the original design perspectives (3).

Let $(d_i)_{i=1}^n \subset \mathcal{D}$ be i.i.d. disturbance patterns generated from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For each d_i, let x_i be the corresponding unknown signal with the associated signature matrix Q_{x_i} as defined in (15). In regard to the average perspective AP, we propose the two-stage (random) optimization program

$$
(24a) \qquad \widetilde{\mathrm{AP}}_1 : \begin{cases} \min\limits_{\gamma, \bar{N}} & \gamma \\ \text{s.t.} & \bar{N}\bar{H} & = 0 \\ & \left\|\bar{N}\bar{F}\right\|_\infty & \geq 1 \\ & \frac{1}{n}\sum\limits_{i=1}^n J\left(\sqrt{\bar{N}Q_{x_i}\bar{N}^\intercal}\right) & \leq \gamma \end{cases}
$$

$$
(24b) \qquad \widetilde{\mathrm{AP}}_2 : \begin{cases} \max\limits_{\bar{N}} & \left\|\bar{N}\bar{F}\right\|_\infty \\ \text{s.t.} & \bar{N}\bar{H} & = 0 \\ & \left\|\bar{N}\right\|_\infty & \leq 1 \\ & \frac{1}{n}\sum\limits_{i=1}^n J\left(\|\bar{N}_1^\star\|_\infty \sqrt{\bar{N}Q_{x_i}\bar{N}^\intercal}\right) & \leq \gamma_1^\star \end{cases}
$$

where $J : \mathbb{R}_+ \to \mathbb{R}_+$ is an increasing and convex payoff function, and in the second stage (24b) $\bar{N}_1^\star$ and $\gamma_1^\star$ are the optimizers of the first stage (24a), i.e., the programs (24) need to be

---

[1] A conservative but easy-to-implement approach is to set G an identity matrix with dimension $k + 1$.

solved sequentially in a *lexicographic* (multi-objective) sense [MA04]. Let us recall that the filter coefficients can always be normalized with no performance deterioration. Hence, it is straightforward to observe that the main goal of the second stage is only to improve the coefficients of $\bar{N}\bar{F}$ (concerning the fault sensitivity) while the optimality of the first stage (concerning the robustification to nonlinearity signatures) is guaranteed. Similarly, we also propose the following two-stage program for the perspective CP:

$$(25a) \qquad \widetilde{\mathrm{CP}}_1 : \begin{cases} \min_{\gamma,\bar{N}} & \gamma \\ \text{s.t.} & \bar{N}\bar{H} & = 0 \\ & \left\| \bar{N}\bar{F} \right\|_\infty & \geq 1 \\ & \max_{i \leq n} \bar{N}Q_{x_i}\bar{N}^\intercal & \leq \gamma \end{cases}$$

$$(25b) \qquad \widetilde{\mathrm{CP}}_2 : \begin{cases} \max_{\bar{N}} & \left\| \bar{N}\bar{F} \right\|_\infty \\ \text{s.t.} & \bar{N}\bar{H} & = 0 \\ & \left\| \bar{N} \right\|_\infty & \leq 1 \\ & \left\| \bar{N}_1^\star \right\|_\infty^2 \left(\max_{i \leq n} \bar{N}Q_{x_i}\bar{N}^\intercal \right) & \leq \gamma_1^\star \end{cases}$$

Remark 4.9 (Computational Complexity). *In view of Lemma 4.3, all the programs in* (24) *and* (25) *can be written as families of convex programs, and hence are tractable. It is, however, worth noting that in case the payoff function of* $\widetilde{\mathrm{AP}}$ *is* $J(\alpha) := \alpha^2$, *the computational complexity of the resulting programs in* (24) *is independent of the number of scenarios* n, *since the problems effectively reduce to a quadratic programming with a constraint involving the average of all the respective signature matrices (i.e.,* $\frac{1}{n}\sum_{i=1}^n Q_{x_i}$). *This is particularly of interest if one requires to train the filter for a large number of scenarios.*

Clearly, the filter designed by programs (24) and (25) is robustified to only finitely many most likely events, and as such, it may remain sensitive to disturbance patterns which have not been observed in the training phase. However, thanks to the probabilistic guarantees detailed in the sequel, we shall show that the probability of such failures (*false alarm*) is low. In fact, the tractability of our proposed scheme comes at the price of allowing for rare threshold violation of the filter. The rest of the subsection formalizes this probabilistic bridge between the program (24) (resp. (25)) and the original perspective AP (resp. CP) in (3) when the class of filters is confined to the linear residuals characterized in Lemma 4.2. For this purpose, we need a technical measurability assumption which is always expected to hold in practice.

Assumption 4.10 (Measurability). *We assume that the mapping* $\mathcal{D} \ni d \mapsto x \in \mathcal{W}_T^{n_x}$ *is measurable where the function spaces are endowed with the* $\mathcal{L}_2$-*topology and the respective Borel sigma-algebra. In particular,* x *can be viewed as a random variable on the same probability space as* d.

Assumption 4.10 is referred to the behavior of the system dynamics as a mapping from the disturbance d to the internal states. In the context of ODEs (6), it is well-known that under mild assumptions (e.g., Lipschitz continuity of E_X) the mapping $d \mapsto X$ is indeed continuous [Kha92, Chapter 5], which readily ensures Assumption 4.10.

4.3.1. *Probabilistic performance of* $\widetilde{\mathrm{AP}}$. Here we study the asymptotic behavior of the empirical average of $\mathbb{E}\big[J(\|r\|)\big]$ uniformly in the filter coefficients $\bar{N}$, which allows us to link the solutions

of programs (24) to AP. Let $\mathcal{N} := \left\{ \bar{N} \in \mathbb{R}^{nr(d_N+1)} : \|\bar{N}\|_\infty \leq 1 \right\}$ and consider the payoff function of AP in (3) as the mapping $\phi : \mathcal{N} \times \mathcal{W}_T^{n_x} \to \mathbb{R}_+$:

$$(26) \qquad \phi(\bar{N}, x) := J\big(\|r_x\|_{\mathcal{L}_2}\big) = J\big(\|\bar{N}\psi_x\|_{\mathcal{L}_2}\big),$$

where the second equality follows from Lemma 4.5.

Theorem 4.11 (Average Performance). *Suppose Assumption 4.10 holds and the random variable x is almost surely bounded[2]. Then, the mapping $\bar{N} \mapsto \phi(\bar{N}, x)$ is a random function. Moreover, if $(x_i)_{i=1}^n \subset \mathcal{W}_T^{n_x}$ are i.i.d. random variables and e_n is the uniform empirical average error*

$$(27) \qquad e_n := \sup_{\bar{N} \in \mathcal{N}} \left\{ \frac{1}{n} \sum_{i=1}^n \phi(\bar{N}, x_i) - \mathbb{E}\big[\phi(\bar{N}, x)\big] \right\},$$

then,

   (i) *the Strong Law of Large Numbers (SLLN) holds, i.e., $\lim_{n \to \infty} e_n = 0$ almost surely.*

   (ii) *the Uniform Central Limit Theorem (UCLT) holds, i.e., $\sqrt{n}e_n$ converges in law to a Gaussian variable with distribution $N(0, \sigma)$ for some $\sigma \geq 0$.*

Proof. See Appendix I.2 along with required preliminaries. □

The following Corollary is an immediate consequence of the UCLT in Theorem 4.11 (ii).

Corollary 4.12. *Let assumptions of Theorem 4.11 hold, and e_n be the empirical average error (27). For all $\varepsilon > 0$ and $k < \frac{1}{2}$, we have*

$$\lim_{n \to \infty} \mathbb{P}^n\big(n^k e_n \geq \varepsilon\big) = 0,$$

where $\mathbb{P}^n$ denotes the n-fold product probability measure on $\big(\Omega^n, \mathcal{F}^n\big)$.

4.3.2. *Probabilistic performance of* $\widetilde{\mathrm{CP}}$. The formulation CP in (3) is known as chance constrained program which has received increasing attention due to recent developments toward tractable approaches, in particular via the scenario counterpart (cf. $\widetilde{\mathrm{CP}}$ in (4)) in a convex setting [CC06, CG08]. These studies are, however, not directly applicable to our problem due to the non-convexity arising from the constraint $\|\bar{N}\bar{F}\|_\infty \geq 1$. Here, following our recent work [MSL15], we exploit the specific structure of this non-convexity and adapt the scenario approach accordingly.

Let $\big(\bar{N}_n^\star, \gamma_n^\star\big)$ be the optimizer obtained through the two-stage programs (25) where $\bar{N}_n^\star$ is the filter coefficients and $\gamma_n^\star$ represents the filter threshold; n is referred to the number of disturbance patterns. Given the filter $\bar{N}_n^\star$, let us denote the corresponding filter residual due to the signal x by $r_x[\bar{N}_n^\star]$; this is a slight modification of our notation r_x in (13) to specify the filter coefficients. To quantify the filter performance, one may ask for the probability that a new unknown signal x violates the threshold $\gamma_n^\star$ when the FDI filter is set to $\bar{N}_n^\star$ (i.e., the probability that $\big\|r_x[\bar{N}_n^\star]\big\|_{\mathcal{L}_2}^2 > \gamma_n^\star$). In the FDI literature such a violation is known as a false alarm, and from the CP standpoint its occurrence probability is allowed at most to the ε level. In this view the performance of the filter can be quantified by the event

$$(28) \qquad \mathcal{E}\big(\bar{N}_n^\star, \gamma_n^\star\big) := \left\{ \mathbb{P}\Big(\big\|r_x[\bar{N}_n^\star]\big\|_{\mathcal{L}_2}^2 > \gamma_n^\star\Big) > \varepsilon \right\}.$$

---

[2] This assumption may be relaxed in terms of the moments of x, though this will not be pursued further here.

The event (28) accounts for the feasibility of the $\widetilde{\text{CP}}$ solution from the original perspective CP. Note that the measure $\mathbb{P}$ in (28) is referred to x whereas the stochasticity of the event stems from the random solutions $(\bar{N}_n^\star, \gamma_n^*)$.[3]

Theorem 4.13 (Chance Performance). *Suppose Assumption 4.10 holds and $(x_i)_{i=1}^n$ are i.i.d. random variables on $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\bar{N}_n^\star \in \mathbb{R}^{n_r(d_N+1)}$ and $\gamma_n^* \in \mathbb{R}_+$ be the solutions of $\widetilde{\text{CP}}$, and measurable in $\mathcal{F}^n$. Then, the set (28) is $\mathcal{F}^n$-measurable, and for every $\beta \in (0,1)$ and any n such that*

$$n \geq \frac{2}{\varepsilon}\Big(\ln \frac{n_f(d_F + d_N + 1)}{\beta} + n_r(d_N + 1) + 1 \Big),$$

where d_N is the degree of the filter and n_f, n_r, d_F are the system size parameters of (5), we have

$$\mathbb{P}^n\Big(\mathcal{E}\big(\bar{N}_n^\star, \gamma_n^*\big) \Big) < \beta.$$

Proof. See Appendix I.2. $\qquad\qquad\square$

5. Cyber-Physical Security of Power Systems: AGC Case Study

In this section, we illustrate the performance of our theoretical results to detect a cyber intrusion in a two-area power system. Motivated by our earlier studies [MVM+10, MVM+11], we consider the IEEE 118-bus power network equipped with primary and secondary frequency control. While the primary frequency control is implemented locally, the secondary loop, referred also as AGC (Automatic Generation Control), is closed over the SCADA system without human operator intervention. As investigated in [MVM+10], a cyber intrusion in this feedback loop may cause unacceptable frequency deviations and potentially load shedding or generation tripping. If the intrusion is, however, detected on time, one may prevent further damage by disconnecting the AGC. We show how to deploy the methodology developed in earlier sections to construct an FDI filter that uses the available measurements to diagnose an AGC intrusion sufficiently fast, despite the presence of unknown load deviations.

5.1. **Mathematical Model Description.** In this section a multi-machine power system, based only on frequency dynamics, is described [Andb]. The system is arbitrarily divided into two control areas. The generators are equipped with primary frequency control and each area is under AGC which adjusts the generating setpoints of specific generators so as to regulate frequency and maintain the power exchange between the two areas to its scheduled value.

5.1.1. *System description.* We consider a system comprising n buses and g number of generators. Let $G = \{i\}_1^g$ denote the set of generator indices and $A_1 = \{i \in G \mid i \text{ in Area 1}\}$, $A_2 = \{i \in G \mid i \text{ in Area 2}\}$ the sets of generators that belong to Area 1 and Area 2, respectively. Let also

$$L_{tie}^k = \{(i,j) | i,j \text{ edges of a tie line from area } k \text{ to the other areas}\},$$

where a tie line is a line connecting the two independently controlled areas and let also $K = \{1,2\}$ be the set of the indices of the control areas in the system.

Using the classical generator model every synchronous machine is modeled as constant voltage source behind its transient reactance. The dynamic states of the system are the rotor angle δ_i (rad), the rotor electrical frequency f_i (Hz) and the mechanical power (output of the turbine)

---

[3]The measure $\mathbb{P}$ is, with slight abuse of notation, the induced measure via the mapping addressed in Assumption 4.10.

P_{mi} (MW) for each generator $i \in G$. We also have one more state that represents the output of the AGC ΔP_{agc_k} for each control area $k \in K$.

We denote by $E_G \in \mathbb{C}^g$ a vector consisting of the generator internal node voltages $E_{Gi} = |E_{Gi}^0| \angle \delta_i$ for $i \in G$. The phase angle of the generator voltage node is assumed to coincide with the rotor angle δ_i and $|E_{Gi}^0|$ is a constant. The voltages of the rest of the nodes are included in $V_N \in \mathbb{C}^n$, whose entries are $V_{Ni} = |V_{Ni}| \angle \theta_i$ for $i = 1, \ldots, n$. To remove the algebraic constraints that appear due to the Kirchhoff's first law for each node, we retain the internal nodes (behind the transient reactance) of the generators and eliminate the rest of the nodes. This could be achieved only under the assumption of constant impedance loads since in that way they can be included in the network admittance matrix. The node voltages can then be linearly connected to the internal node voltages, and hence to the dynamic state δ_i. This results in a reduced admittance matrix that corresponds only to the internal nodes of the generators, where the power flows are expressed directly in terms of the dynamic states of the system. The resulting model of the two area power system is described by the following set of equations.

$$\dot{\delta}_i = 2\pi(f_i - f_0),$$

$$\dot{f}_i = \frac{f_0}{2 H_i S_{B_i}} (P_{m_i} - P_{e_i}(\delta) - \frac{1}{D_i}(f_i - f_0) - \Delta P_{load_i}),$$

$$\dot{P}_{m,a_k} = \frac{1}{T_{ch,a_k}} (P_{m,a_k}^0 + v_{a_k} \Delta P_{p,a_k}^{sat} + w_{a_k} \Delta P_{agc,k}^{sat} - P_{m,a_k}),$$

$$\Delta \dot{P}_{agc,k} = \sum_{j \in A_k} c_{kj}(f_j - f_0) + \sum_{j \in A_k} b_{kj}(P_{m_j} - P_{e_j}(\delta) - \Delta P_{load_j})$$

$$- \frac{1}{T_{N_k}} g_k(\delta, f) - C_{p_k} h_k(\delta, f) - \frac{K_k}{T_{N_k}}(\Delta P_{agc,k} - \Delta P_{agc,k}^{sat}).$$

where $i \in G$, $a_k \in A_k$ for $k \in K$. Supperscript sat on the AGC output signal $\Delta P_{agc,k}$ and on the primary frequency control signal $\Delta P_{p,a_k}$ highlights the saturation to which the signals are subjected. The primary frequency control is given by $\Delta P_{p,i} = -(f_i - f_0)/S_i$. Based on the reduced admittance matrix, the generator electric power output is given by

$$P_{ei} = \sum_{j=1}^{g} E_{G_i} E_{G_j} (G_{ij}^{red} \cos(\delta_i - \delta_j) + B_{ij}^{red} \sin(\delta_i - \delta_j)).$$

Moreover, $g_k := \sum_{(i,j) \in L_{tie}^k} (P_{ij} - P_{T_{12}^0})$ and $h_k := \frac{dg_k}{dt}$, where the power flow P_{ij}, based on the initial admittance matrix of the system, is given by

$$P_{ij} = |V_{Ni}||V_{Nj}|(G_{ij} \cos(\theta_i - \theta_j) + B_{ij} \sin(\theta_i - \theta_j))$$

All undefined variables are constants, and details on the derivation of the models can be found in [MVAL12]. The AGC attack is modeled as an additive signal to the AGC signal. For instance, if the attack signal is imposed in Area 1, the mechanical power dynamics of Area 1 will be modified as

$$\dot{P}_{m,a_1} = \frac{1}{T_{ch,a_1}} (P_{m,a_1}^0 + v_{a_1} \Delta P_{p,a_1}^{sat} + w_{a_1}(\Delta P_{agc_1}^{sat} + f(t)) - P_{m,a_1}),$$

The above model can be compactly written as

$$(29) \qquad \begin{cases} \dot{X}(t) = h(X(t)) + B_d d(t) + B_f f(t) \\ Y(t) = CX(t), \end{cases}$$
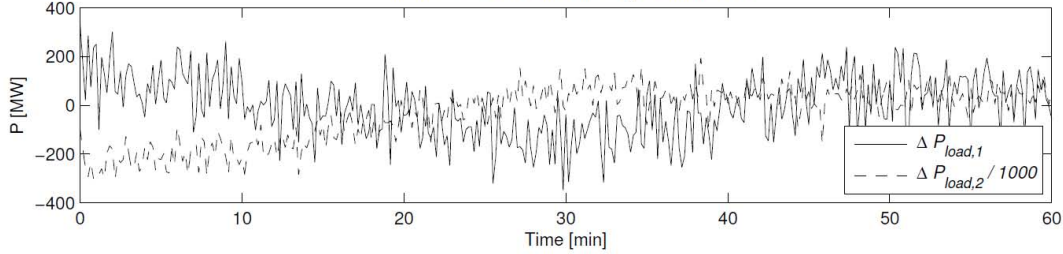
FIGURE 2. Stochastic load fluctuation and prediction error [Anda, p. 59]

where $X := \left[\{\delta_i\}_{1:g}, \{f_i\}_{1:g}, \{P_{m,i}\}_{1:g}, \{\Delta P_{agc_i}\}_{1:2} \right]^\mathsf{T} \in \mathbb{R}^{3g+2}$ denotes the internal states vector comprising rotor angles δ_i, generators frequencies f_i, generated mechanical powers $P_{m,i}$, and the AGC control signal ΔP_{agc_i} for each area. The external input $d := \left[\{\Delta P_{load_i}\}_{1:g} \right]^\mathsf{T}$ represents the unknown load disturbances (discussed in the next subsection), and f represents the intrusion signal injected to the AGC of the first area. We assume that the measurements of all the frequencies and generated mechanical power are available, i.e., $Y = \left[\{f_i\}_{1:g}, \{P_{m,i}\}_{1:g} \right]^\mathsf{T} \in \mathbb{R}^{2g}$. The nonlinear function $h(\cdot)$ and the constant matrices B_d, B_f and C can be easily obtained by the mapping between the analytical model and (29). To transfer the ODE dynamic expression (29) into the DAE (5) it suffices to introduce

$$x := \begin{bmatrix} X - X_e \\ d \end{bmatrix}, \qquad z := Y - CX_e$$

$$E(x) := \begin{bmatrix} h(X) - A(X - X_e) \\ 0 \end{bmatrix}, \quad H(p) := \begin{bmatrix} -pI + A & B_d \\ C & 0 \end{bmatrix}, \quad L(p) := \begin{bmatrix} 0 \\ -I \end{bmatrix}, \quad F(p) := \begin{bmatrix} B_f \\ 0 \end{bmatrix},$$

where X_e is the equilibrium of (29), i.e., $h(X_e) = 0$, and $A := \frac{\partial h}{\partial X}\big|_{X=X_e}$. Notice that by the above definition, the nonlinear term $E(\cdot)$ only carries the nonlinearity of the system while the linear terms of the dynamic are incorporated into the constant matrices H, L, F. This can always be done without loss of generality, and practically may improve the performance of the scheme, as the linear terms can be fully decoupled from the residual of the filter.

5.1.2. *Load Deviations and Disturbances.* Small power imbalances arise during normal operation of power networks due, for example, to load fluctuation, load forecast errors, and trading on electricity market. Each of these sources give rise to deviations at different time scale. High frequency load fluctuation is typically time uncorrelated stochastic noise on a second or minute time scale, whereas forecast errors usually stem from the mismatch of predicted and actual consumption on a 15-minute time scale. Figure 2 demonstrates two samples of stochastic load fluctuation and forecast error which may appear at two different nodes of the network [Anda, p. 59]. The trading on the electricity market also introduces disturbances, for example, in an hourly framework (depending on the market).

To capture these sources of uncertainty we consider a space of disturbance patterns comprising combinations of sinusoids at different frequency ranges (to model short term load fluctuation and mid-term forecast errors) and step functions (to model long-term abrupt changes due to the market). The space of load deviations (i.e., the disturbance patterns $\mathcal{D}$ in our FDI setting) is

then described by

$$(30) \qquad \Delta P_{load}(t) := \alpha_0 + \sum_{i=1}^{\eta} \alpha_i \sin(\omega_i t + \phi_i), \qquad t \in [0, T],$$

where the parameters $(\alpha_i)_{i=0}^{\eta}$, $(\omega_i)_{i=1}^{\eta}$ $(\phi_i)_{i=1}^{\eta}$, and η are random variables whose distributions induce the probability measure on $\mathcal{D}$. We assume that $\sum_{i=0}^{\eta} |\alpha_i|^2$ is uniformly bounded with probability 1 to meet the requirements of Theorem 4.11.

5.2. **Diagnosis Filter Design.** To design the FDI filter, we set the degree of the filter $d_N = 7$, the denominator $a(p) = (p + 2)^{d_N}$, and the finite time horizon $T = 10 \sec$. Note that the degree of the filter is significantly less than the dimension of the system (29), which is 59. This is a general advantage of the residual generator approach in comparison to the observer-based approach where the filter order is effectively the same as the system dynamics. To compute the signature matrix Q_x, we resort to the finite dimensional approximation $Q_{\mathcal{B}}$ in Proposition 4.7. Inspired by the class of disturbances in (30), we first choose Fourier basis with 80 harmonics

$$(31) \qquad b_i(t) := \begin{cases} \cos(\frac{i}{2}\omega t) & i : \text{even} \\ \sin(\frac{i+1}{2}\omega t) & i : \text{odd} \end{cases}, \qquad \omega := \frac{2\pi}{T}, \quad i \in \{0, 1, \cdots, 80\}.$$

We should emphasize that there is no restriction on the basis selection as long as Assumptions 4.6 are fulfilled; we refer to [MVAL12, Section V.B] for another example with a polynomial basis. Given the basis (31), it is easy to see that the differentiation matrix D introduced in (19) is

$$D = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \omega & \cdots & 0 & 0 \\ 0 & -\omega & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & & 0 & 80\omega \\ 0 & 0 & 0 & \cdots & -80\omega & 0 \end{bmatrix}.$$

We can also compute offline (independent of x) the matrix G in (21) with the help of the basis (31) and the denominator $a(p)$. To proceed with Q_x of a sample ΔP_{load} we need to run the system dynamic (29) with the input $d(\cdot) := \Delta P_{load}$ and compute $x(t) := [X(t)^\intercal, \Delta P_{load}(t)]^\intercal$ where X is the internal states of the system. Given the signal x, we then project the nonlinearity signature $t \mapsto e_x(t) =: E(x(t))$ onto the subspace $\mathcal{B}$ (i.e., $\mathbb{T}_{\mathcal{B}}(e_x)$), and finally obtain Q_x from (22). In the following simulations, we deploy the YALMIP toolbox [Lof04] to solve the corresponding optimization problems.

5.3. **Simulation Results.**

5.3.1. *Test system.* To illustrate the FDI methodology we employed the IEEE 118-bus system. The data of the model are retrieved from a snapshot available at [ref]. It includes 19 generators, 177 lines, 99 load buses and 7 transmission level transformers. Since there were no dynamic data available, typical values provided by [AF02] were used for the simulations. The network was arbitrarily divided into two control areas whose nonlinear frequency model was developed in the preceding subsections. Figure 3 depicts a single-line diagram of the network and the boundaries of the two controlled areas where the first and second area contain, respectively, 12 and 7 generators.
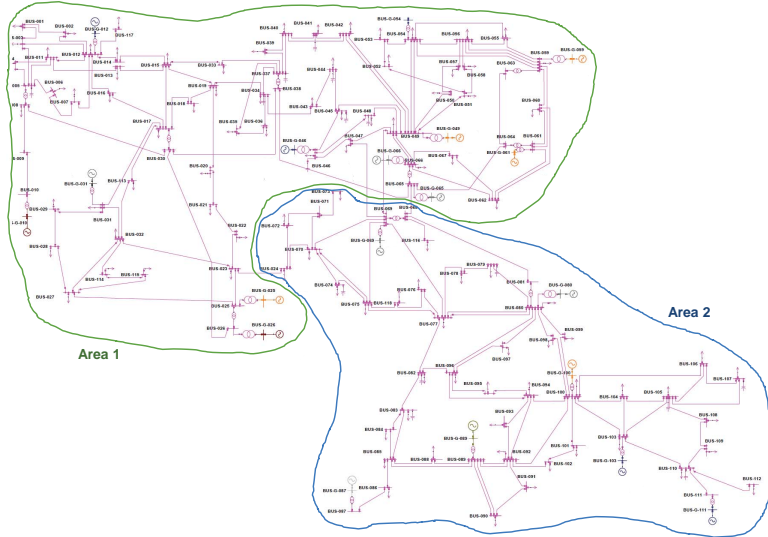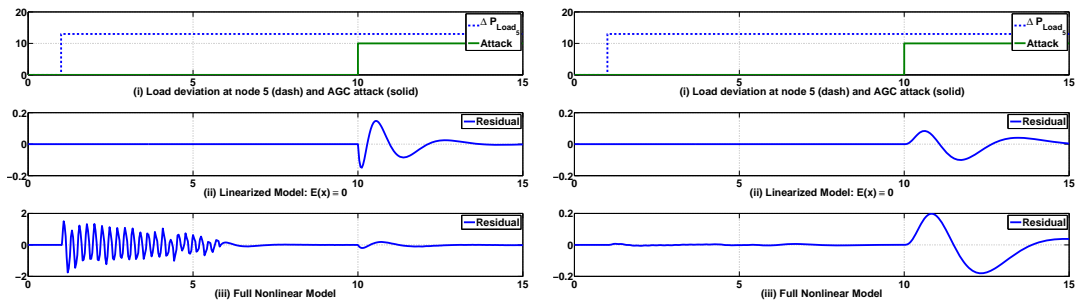
FIGURE 3. IEEE 118-bus system divided into two control areas

(a) Performance of the filter neglecting the nonlinear term

(b) Performance of the filter trained for the step signatures

FIGURE 4. Performance of the FDI filters with step inputs

5.3.2. *Numerical results.* In the first simulation we consider the scenario that an attacker manipulates the AGC signal of the first area at $T_{ack} = 10 \sec$. We model this intrusion as a step signal equal to 14 MW injected into the AGC in Area 1. To challenge the filter, we also assume that a step load deviation occurs at $T_{load} = 1 \sec$ at node 5. In the following we present the results of two filters: Figure 4(a) shows the filter based on formulation (12) in Approach (I), which basically neglects the nonlinear term; Figure 4(b) shows the proposed filter in (24) based on AP perspective where the payoff function is $J(\alpha) := \alpha^2$; see Remark 4.9 why such a payoff function is of particular interest.

We validate the filters performance with two sets of measurements: first the measurements obtained from the linearized dynamic (i.e. $E(x) \equiv 0$); second the measurements obtained from the full nonlinear model (29). As shown in Fig. 4(a)(ii) and Fig. 4(b)(ii), both filters work perfectly well with linear dynamics measurements. It even appears that the first filter seems
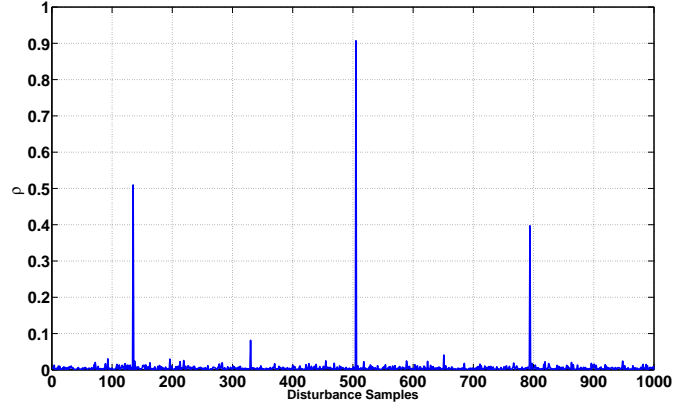
FIGURE 5. The indicator ρ defined in (32)

more sensitive. However, Fig. 4(a)(iii) and Fig. 4(b)(iii) demonstrate that in the nonlinear setting the first filter fails whereas the robustified filter works effectively similar to the linear setting.

In the second simulation, to evaluate the filter performance in more realistic setup, we robustify the filter to random disturbance patterns, and then verify it with new generated samples. To measure the performance in the presence of the attack, we introduce the following indicator:

$$(32) \qquad \rho := \frac{\max\limits_{t \leq T_{ack}} \|r(t)\|_\infty}{\max\limits_{t \leq T} \|r(t)\|_\infty},$$

where r is the residual (11), and T_{ack} is when the attack starts. Observe that $\rho \in [0, 1]$, and the lower ρ the better performance for the filter, e.g., in Fig. 4(a)(iii) $\rho = 1$, and in Fig. 4(b)(iii) $\rho \approx 0$.

In the training phase, we randomly generate five sinusoidal load deviations as described in (30), and excite the dynamics for $T = 10\ sec$ in the presence of each of the load deviations individually. Hence, in total we have $n = 19 \times 5 = 95$ disturbance signatures. Then, we compute the filter coefficients by virtue of $\widetilde{\text{AP}}$ in (24) with the payoff function $J(\alpha) := \alpha^2$ and these 95 samples. In the operation phase, we generate two new disturbance patterns with the same distribution as in the training phase and run the system in the presence of both load deviations simultaneously at two random nodes for the horizon $T = 120\ sec$. Meanwhile, we inject an attack signal at $T_{ack} = 110\ sec$ in the AGC, and compute the indicator ρ in (32). Figure 5 demonstrates the result of this simulation for 1000 experiments.

6. Conclusion and Future Directions

In this article, we proposed a novel perspective toward the FDI filter design, which is tackled via an optimization-based methodology along with probabilistic performance guarantees. Thanks to the convex formulation, the methodology is applicable to high dimensional nonlinear systems in which some statistical information of exogenous disturbances are available. Motivated by our earlier works, we deployed the proposed technique to design a diagnosis filter to detect the AGC malfunction in two-area power network. The simulation results validated the filter performance, particularly when the disturbance patterns are different from training to the operation phase.

The central focus of the work here is to robustify the filter to certain signatures of dynamic nonlinearities in the presence of given disturbance patterns. As a next step, motivated by applications that the disruptive attack may follow certain patterns, a natural question is whether the filter can be trained to these attack patterns. From the technical standpoint, this problem in principle may be different from the robustification process since the former may involve maximization of the residual norm as opposed to the minimization for the robustification discussed in this article. Therefore, this problem offers a challenge to reconcile the disturbance rejection and the fault sensitivity objectives.

The proposed methodology in this study is applicable to both discrete and continuous-time dynamics and measurements. In reality, however, we often have different time-setting in different parts, i.e., we only have discrete-time measurements while the system dynamics follows a continuous-time behavior. We believe this setup introduces new challenges to the field. We recently reported heuristic attempts toward this objective in [ETML13], though there is still a need to address this problem in a rigorous and systematic framework.

I. Appendix

I.1. **Proofs of Section 4.2.** Let us start with a preliminary required for the main proof of this section.

Lemma I.1. *Let* $N(p) := \sum_{i=0}^{d_N} N_i p^i$ *be an* $\mathbb{R}^{n_r}$ *row polynomial vector with degree* d_N, *and* $a(p)$ *be a stable polynomial with the degree at least* d_N. *Let* $\bar{N} := [N_0 \ N_1 \ \cdots \ N_{d_N}]$ *be the collection of the coefficients of* $N(p)$. *Then,*

$$\left\| a^{-1} N \right\|_{\mathcal{H}_\infty} \leq \widetilde{C} \|\bar{N}\|_\infty, \qquad \widetilde{C} := \sqrt{n_r (d_N + 1)} \ \|a^{-1}\|_{\mathcal{H}_\infty}.$$

Proof. Let $b(p) := \sum_{i=0}^{d_N} b_i p^i$ be a polynomial scaler function. By $\mathcal{H}_\infty$-norm definition we have

$$(\text{I.1}) \qquad \left\| a^{-1} b \right\|_{\mathcal{H}_\infty}^2 = \sup_{\omega \in (-\infty, \infty)} \left| \frac{b(j\omega)}{a(j\omega)} \right|^2 \leq \sup_{\omega \in [0, \infty)} \frac{\sum_{i=0}^{d_N} |b_i|^2 \omega^{2i}}{|a(j\omega)|^2}.$$

Let $\bar{b} := \begin{bmatrix} b_0 & b_1 & \cdots & b_{d_N} \end{bmatrix}$. It is then straightforward to inspect that

$$(\text{I.2}) \qquad \sum_{i=0}^{d_N} |b_i|^2 \omega^{2i} \leq \begin{cases} (d_N + 1)\|\bar{b}\|_\infty^2 & \text{if} \quad \omega \in [0, 1] \\ (d_N + 1)\|\bar{b}\|_\infty^2 \omega^{2d_N} & \text{if} \quad \omega \in (1, \infty) \end{cases}$$

Therefore, (I.1) together with (I.2) yields to

$$\left\| a^{-1} b \right\|_{\mathcal{H}_\infty}^2 \leq (d_N + 1) \left\| a^{-1} \right\|_{\mathcal{H}_\infty}^2 \|\bar{b}\|_\infty^2.$$

Now, taking the dimension of the vector $N(p)$ into consideration, we conclude the desired assertion. $\square$

Proof of Lemma 4.5. Let $\ell \geq d_N$ be the degree of the scalar polynomial $a(p)$. Then, taking advantage of the state-space representation of the matrix transfer function $a^{-1}(p)N(p)$, in particular the observable canonical form [ZD97, Section 3.5], we have

$$r_x(t) = \int_0^t C\,\mathrm{e}^{-A(t-\tau)}\,Be_x(\tau)\mathrm{d}\tau + De_x(t),$$

where $C \in \mathbb{R}^{1\times\ell}$ is a constant vector, $A \in \mathbb{R}^{\ell\times\ell}$ is the state matrix depending only on $a(p)$, and $B \in \mathbb{R}^{\ell\times n_r}$ and $D \in \mathbb{R}^{1\times n_r}$ are matrices that depend linearly on all the coefficients of the numerator $\bar{N} \in \mathbb{R}^{n_r(d_N+1)}$. Therefore, it can be readily deduced that (14a) holds for some function $\psi_x \in \mathcal{W}_T^{n_r(d_N+1)}$. In regard to (14a) and the definition (13), we have

$$(\mathrm{I.3}) \qquad \|\bar{N}\psi_x\|_{\mathcal{L}_2} = \|r_x\|_{\mathcal{L}_2} = \left\|a^{-1}(p)N(p)e_x\right\|_{\mathcal{L}_2} \leq \left\|a^{-1}N\right\|_{\mathcal{H}_\infty}\|e_x\|_{\mathcal{L}_2} \leq \widetilde{C}\|\bar{N}\|_\infty\|e_x\|_{\mathcal{L}_2},$$

where the first inequality follows from the classical result that the $\mathcal{L}_2$-gain of a matrix transfer function is the $\mathcal{H}_\infty$-norm of the matrix [ZD97, Theorem 4.3, p. 51], and the second inequality follows from Lemma I.1. Since (I.3) holds for every $\bar{N} \in \mathbb{R}^{n_r(d_N+1)}$, then

$$\|\psi_x\|_{\mathcal{L}_2} \leq \sqrt{n_r(d_N+1)}\,\widetilde{C}\|e_x\|_{\mathcal{L}_2},$$

which implies (14b). □

Proof of Proposition 4.7. Observe that by virtue of the triangle inequality and linearity of the projection mapping we have

$$\left|\|r_x\|_{\mathcal{L}_2} - \left\|a^{-1}(p)N(p)\mathbb{T}_\mathcal{B}(e_x)\right\|_{\mathcal{L}_2}\right| \leq \left\|a^{-1}(p)N(p)\big(e_x - \mathbb{T}_\mathcal{B}(e_x)\big)\right\|_{\mathcal{L}_2} \leq \widetilde{C}\|\bar{N}\|_\infty\delta,$$

where the second inequality follows in the same spirit as (I.3) and $\|e_x - \mathbb{T}_\mathcal{B}(e_x)\|_{\mathcal{L}_2} \leq \delta$. Note that by definitions of Q_x and $Q_\mathcal{B}$ in (15) and (22), respectively, we have

$$\left|\bar{N}(Q_x - Q_\mathcal{B})\bar{N}^\intercal\right| = \left|\|r_x\|_{\mathcal{L}_2}^2 - \left\|a^{-1}(p)N(p)\mathbb{T}_\mathcal{B}(e_x)\right\|_{\mathcal{L}_2}^2\right| \leq \widetilde{C}\|\bar{N}\|_\infty\delta\big(\widetilde{C}\|\bar{N}\|_\infty\delta + 2\|r_x\|_{\mathcal{L}_2}\big)$$
$$\leq \widetilde{C}^2\|\bar{N}\|_\infty^2\delta\big(\delta + 2\|e_x\|_{\mathcal{L}_2}\big) \leq C\|a^{-1}\|_{\mathcal{H}_\infty}\|\bar{N}\|_2^2\delta\big(1 + 2\|e_x\|_{\mathcal{L}_2}\big)$$

where the inequality of the first line stems from the simple inequality $|\alpha^2 - \beta^2| \leq |\alpha - \beta|(2|\alpha| + |\alpha - \beta|)$, and C is the constant as in (14b). □

I.2. **Proofs of Section 4.3.** To prove Theorem 4.11 we need a preparatory result addressing the continuity of the mapping ϕ in (26).

Lemma I.2. *Consider the function ϕ as defined in (26). Then, there exists a constant $L > 0$ such that for any $\bar{N}_1, \bar{N}_2 \in \mathcal{N}$ and $x_1, x_2 \in \mathcal{W}_T^{n_x}$ where $\|x_i\|_{\mathcal{L}_2} \leq M$, we have*

$$\left|\phi(\bar{N}_1, x_1) - \phi(\bar{N}_2, x_2)\right| \leq L\big(\left\|\bar{N}_1 - \bar{N}_2\right\|_\infty + \|x_1 - x_2\|_{\mathcal{L}_2}\big).$$

Proof. Let L_E be the Lipschitz continuity constant of the mapping $E : \mathbb{R}^{n_x} \to \mathbb{R}^{n_r}$ in (13). We modify the notation of r_x in (13) with a new argument as $r_x[\bar{N}]$, in which $\bar{N}$ represents the filter coefficients. Then, with the aid of (I.3), we have

$$\sup_{\|x\|_{\mathcal{L}_2}\leq M}\sup_{\bar{N}\in\mathcal{N}}\|r_x[\bar{N}]\|_{\mathcal{L}_2} \leq \sup_{\|x\|_{\mathcal{L}_2}\leq M}\sup_{\bar{N}\in\mathcal{N}}\widetilde{C}L_E\|\bar{N}\|_\infty\|x\|_{\mathcal{L}_2} \leq \widetilde{M}, \qquad \widetilde{M} \coloneqq \widetilde{C}L_E M,$$

where the constant $\widetilde{C}$ is introduced in Lemma I.1. As the payoff function J is convex, it is then Lipschitz continuous over the compact set $[0, \widetilde{M}]$ [Ber09, Proposition 5.4.2, p. 185]; we denote this Lipschitz constant by L_J. Then for any $\bar{N}_i \in \mathcal{N}$ and $\|x_i\|_{\mathcal{L}_2} \leq M$, $i \in \{1, 2\}$, we have,

$$\left|\phi(\bar{N}_1, x_1) - \phi(\bar{N}_2, x_2)\right| \leq L_J\left|\left\|r_{x_1}[N_1]\right\|_{\mathcal{L}_2} - \left\|r_{x_2}[N_2]\right\|_{\mathcal{L}_2}\right|$$

$$\le L_J\Big(\big\|r_{x_1}[N_1] - r_{x_1}[N_2]\big\|_{\mathcal{L}_2} + \big\|r_{x_1}[N_2] - r_{x_2}[N_2]\big\|_{\mathcal{L}_2}\Big)$$

(I.4)

$$\le L_J\Big(\widetilde{C}\|e_{x_1}\|_{\mathcal{L}_2}\|N_1 - N_2\|_\infty + \widetilde{C}\|e_{x_1} - e_{x_2}\|_{\mathcal{L}_2}\|N_2\|_\infty\Big)$$

$$\le L_J\widetilde{C}L_E\big(M\|N_1 - N_2\|_\infty + \|x_1 - x_2\|_{\mathcal{L}_2}\big).$$

where (I.4) follows from (I.3) and the fact that the mapping $(\bar{N}, e_x) \mapsto r_x[\bar{N}]$ is bilinear. $\qquad\square$

Proof of Theorem 4.11. By virtue of Lemma I.2, one can infer that for every $\bar{N} \in \mathcal{N}$ the mapping $x \mapsto \phi(\bar{N}, x)$ is continuous, and hence measurable. Therefore, $\phi(\bar{N}, x)$ can be viewed as a random variable for each $\bar{N} \in \mathcal{N}$, which yields to the first assertion, see [Bil99, Chapter 2, p. 84] for more details.

By uniform (almost sure) boundedness and again Lemma I.2, the mapping $\bar{N} \mapsto \phi(\bar{N}, x)$ is uniformly Lipschitz continuous (except on a negligible set), and consequently first moment continuous in the sense of [Han12, Definition 2.5]. We then reach (i) by invoking [Han12, Theorem 2.1].

For assertion (ii), note that the compact set $\mathcal{N}$ is finite dimensional, and thus admits a logarithmic ε-capacity in the sense of [Dud99, Section. 1.2, p. 11]. Therefore, the condition [Dud99, (6.3.4), p. 209] is satisfied. Since the other requirements of [Dud99, Theorem 6.3.3, p. 208] are readily fulfilled by the uniform boundedness assumption and Lemma I.2, we arrive at the desired UCLT assertion in (ii). $\qquad\square$

To keep the paper self-contained, we provide a proof for Theorem 4.13 in the following, but refer the interested reader to [MSL15, Theorem 4.1] for a result of a more general setting.

Proof of Theorem 4.13. The measurability of $\mathcal{E}$ is a straightforward consequence of the measurability of $[\bar{N}_n^\star, \gamma_n^\star]$ and Fubini's Theorem [Bil95, Theorem 18.3, p. 234]. For notational simplicity, we introduce the following notation. Let $\ell := n_r(d_N + 1) + 1$ and define the function $f : \mathbb{R}^\ell \times \mathcal{W}_T^{n_x} \to \mathbb{R}$

$$f(\theta, x) := \bar{N}Q_x\bar{N}^\intercal - \gamma, \qquad \theta := [\bar{N}, \gamma]^\intercal \in \mathbb{R}^\ell,$$

where Q_x is the nonlinearity signature matrix of x as defined in (15), and θ is the augmented vector collecting all the decision variables. Consider the convex sets $\Theta_j \subset \mathbb{R}^\ell$

$$\Theta_j := \Big\{\theta = [\bar{N}, \gamma]^\intercal \mid \bar{N}\bar{H} = 0, \ \bar{N}\bar{F}v_j \ge 1\Big\}, \qquad v_j := [0, \cdots, \overset{\downarrow\ j^{\text{th}}}{1}, \cdots, 0]^\intercal,$$

where the size of v_j is $m := n_f(d_F + d_N + 1)$. Note that in view of Lemma 4.3, we can replace the characterization of the filter coefficients in (10) with $\theta \in \bigcup_{j=1}^m \Theta_j$. We then express the program CP in (3) and its random counterpart $\widetilde{\mathrm{CP}}_1$ in (25a) as follows:

$$\mathrm{CP} : \begin{cases} \displaystyle\min_{\theta \in \bigcup_{j=1}^m \Theta_j} \ c^\intercal\theta \\ \text{s.t.} \quad \mathbb{P}\big(f(\theta, x) \le 0\big) \ge 1 - \varepsilon \end{cases} \qquad \widetilde{\mathrm{CP}}_1 : \begin{cases} \displaystyle\min_{\theta \in \bigcup_{j=1}^m \Theta_j} \ c^\intercal\theta \\ \text{s.t.} \quad \displaystyle\max_{i \le n} f(\theta, x_i) \le 0, \end{cases}$$

where c is the constant vector with 0 elements except the last which is 1. It is straightforward to observe that the optimal threshold $\gamma_n^\star$ of the two-stage program $\widetilde{\mathrm{CP}}$ in (25) is the same as the optimal threshold obtained in the first stage $\widetilde{\mathrm{CP}}_1$. Thus, it suffices to show the desired assertion

considering only the first stage. Let $\theta_n^\star := [\bar{N}_n^\star, \gamma_n^\star]$ denote the optimizer of $\widetilde{\mathrm{CP}}_1$. Now, consider m sub-programs denoted by $\mathrm{CP}(j)$ and $\widetilde{\mathrm{CP}}(j)$ for $j \in \{1, \cdots, m\}$:

$$\mathrm{CP}(j) : \begin{cases} \min_{\theta \in \Theta_j} & c^\mathsf{T}\theta \\ \text{s.t.} & \mathbb{P}\big(f(\theta, x) \leq 0\big) \geq 1 - \varepsilon \end{cases} \qquad \widetilde{\mathrm{CP}}(j) : \begin{cases} \min_{\theta \in \Theta_j} & c^\mathsf{T}\theta \\ \text{s.t.} & \max_{i \leq n} f(\theta, x_i) \leq 0, \end{cases}$$

Let us denote the optimal solution of $\widetilde{\mathrm{CP}}(j)$ by $\theta_{n,j}^*$. Note that for all j, the set Θ_j is deterministic (not affected by x) and convex, and the corresponding random program $\widetilde{\mathrm{CP}}(j)$ is feasible if $\Theta_j \neq \emptyset$, thanks to the min-max structure of $\widetilde{\mathrm{CP}}(j)$. Therefore, we can readily employ the existing results of the random convex problems. Namely, by [CG08, Theorem 1] we have

$$\mathbb{P}^n\big(\mathcal{E}(\theta_{n,j}^*)\big) < \sum_{i=0}^{\ell-1} \binom{n}{i} \varepsilon^i (1-\varepsilon)^{n-i}, \qquad \forall j \in \{1, \cdots, m\}$$

where $\mathcal{E}$ is introduced in (28). Furthermore, it is not hard to inspect that $\theta_n^* \in \big(\theta_{n,j}^*\big)_{j=1}^m$. Thus, $\mathcal{E}(\theta_n^*) \subseteq \bigcup_{j=1}^m \mathcal{E}(\theta_{n,j}^*)$ which yields

$$\mathbb{P}^n\big(\mathcal{E}(\theta_n^*)\big) \leq \mathbb{P}^n\Big(\bigcup_{j=1}^m \mathcal{E}(\theta_{n,j}^*)\Big) \leq \sum_{j=1}^m \mathbb{P}^n\big(\mathcal{E}(\theta_{n,j}^*)\big) < m \sum_{i=0}^{\ell-1} \binom{n}{i} \varepsilon^i (1-\varepsilon)^{n-i}.$$

Now, considering β as an upper bound, the desired assertion can be obtained by similar calculation as in [Cal09] to make the above inequality explicit for n in terms of ε and β. $\qquad \square$

References

[Ada75] Robert A. Adams, *Sobolev spaces*, Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975, Pure and Applied Mathematics, Vol. 65.

[AF02] P. M. Anderson and A. A. Fouad, *Power System Control and Stability*, IEEE Computer Society Press, 2002.

[Anda] Gran Andersson, *Dynamics and control of electric power systems*, Power System Laboratory, ETH Zurich,.

[Andb] ———, *Power system analysis*, Power System Laboratory, ETH Zurich,.

[Bea71] R. V. Beard, *Failure accommodation in linear systems through self-reorganization*, Ph.D. thesis, Massachusetts Inst. Technol., Cambridge, MA, 1971.

[Ber09] Dimitri P. Bertsekas, *Convex Optimization Theory*, Athena Scientific, 2009. MR 2830150 (2012f:90001)

[Bil95] Patrick Billingsley, *Probability and measure*, third ed., Wiley, 1995. MR 1324786 (95k:60001)

[Bil99] ———, *Convergence of Probability Measures*, second ed., Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons Inc., New York, 1999. MR 1700749 (2000e:60008)

[Cal09] Giuseppe C. Calafiore, *A note on the expected probability of constraint violation in sampled convex programs*, 18th IEEE International Conference on Control Applications Part of 2009 IEEE Multi-conference on Systems and Control, july 2009, pp. 1788 – 1791.

[CC06] Giuseppe C. Calafiore and Marco C. Campi, *The scenario approach to robust control design*, IEEE Trans. Automat. Control **51** (2006), no. 5, 742–753. MR 2232597 (2007a:93075)

[CG08] M. C. Campi and S. Garatti, *The exact feasibility of randomized solutions of uncertain convex programs*, SIAM J. Optim. **19** (2008), no. 3, 1211–1230. MR 2460739 (2009j:90081)

[CP82] J. Chen and R. Patton, *Robust model based faults diagnosis for dynamic systems*, Dordrecht: Kluwer Academic Publishers, New York, 1982.

[CS98] Walter H. Chung and Jason L. Speyer, *A game-theoretic fault detection filter*, IEEE Trans. Automat. Control **43** (1998), no. 2, 143–161.

[DS99] Randal K Douglas and Jason L Speyer, *H bounded fault detection filter*, Journal of guidance, control, and dynamics **22** (1999), no. 1, 129–138.

[Dud99]   R. M. Dudley, *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics, vol. 63, Cambridge University Press, Cambridge, 1999. MR 1720712 (2000k:60050)

[EFK13]   Daniel Eriksson, Erik Frisk, and Mattias Krysander, *A method for quantitative fault diagnosability analysis of stochastic linear descriptor models*, Automatica **49** (2013), no. 6, 1591–1600.

[ETML13]   Erasmia Evangelia Tiniou, Peyman Mohajerin Esfahani, and John Lygeros, *Fault detection with discrete-time measurements: an application for the cyber security of power networks*, 52nd IEEE Conference Decision and Control, Dec 2013, pp. 194–199.

[FF12]   Giuseppe Franzè and Domenico Famularo, *A robust fault detection filter for polynomial nonlinear systems via sum-of-squares decompositions*, Systems & Control Letters **61** (2012), no. 8, 839–848.

[FKA09]   Erik Frisk, Mattias Krysander, and Jan Aslund, *Sensor placement for fault isolation in linear differential-algebraic systems*, Automatica **45** (2009), no. 6, 364–371.

[Han12]   Lars Peter Hansen, *Proofs for large sample properties of generalized method of moments estimators*, Journal of Econometrics **170** (2012), no. 2, 325–330. MR 2970318

[HKEY99]   H. Hammouri, M. Kinnaert, and E.H. El Yaagoubi, *Observer-based approach to fault detection and isolation for nonlinear systems*, Automatic Control, IEEE Transactions on **44** (1999), no. 10, 1879–1884.

[HP96]   M. Hou and R.J. Patton, *An lmi approach to H_-/H_∞ fault detection observers*, Control '96, UKACC International Conference on (Conf. Publ. No. 427), vol. 1, sept. 1996, pp. 305 – 310 vol.1.

[Jon73]   H. L. Jones, *Failure detection in linear systems*, Ph.D. thesis, Massachusetts Inst. Technol., Cambridge, MA, 1973.

[Kha92]   Hassan K. Khalil, *Nonlinear systems*, Macmillan Publishing Company, New York, 1992. MR 1201326 (93k:34001)

[Lof04]   J. Lofberg, *Yalmip : a toolbox for modeling and optimization in matlab*, Computer Aided Control Systems Design, 2004 IEEE International Symposium on, sept. 2004, pp. 284 –289.

[Lue69]   David G. Luenberger, *Optimization by vector space methods*, John Wiley & Sons Inc., New York, 1969.

[MA04]   R Timothy Marler and Jasbir S Arora, *Survey of multi-objective optimization methods for engineering*, Structural and multidisciplinary optimization **26** (2004), no. 6, 369–395.

[Mas86]   Mohammad-Ali Massoumnia, *A geometric approach to the synthesis of failure detection filters*, IEEE Trans. Automat. Control **31** (1986), no. 9, 839–846.

[MEVAL]   Peyman Mohajerin Esfahani, Maria Vrakopoulou, Goran Andersson, and John Lygeros, *Intrusion detection in electric power networks*, Patent applied for EP-12005375, filed 24 July 2012.

[MSL15]   Peyman Mohajerin Esfahani, Tobias Sutter, and John Lygeros, *Performance bounds for the scenario approach and an extension to a class of non-convex programs*, IEEE Transactions on Automatic Control **60** (2015), no. 1, 46–58.

[MVAL12]   Peyman Mohajerin Esfahani, Maria Vrakopoulou, Goran Andersson, and John Lygeros, *A tractable nonlinear fault detection and isolation technique with application to the cyber-physical security of power systems*, 51st IEEE Conference Decision and Control, Dec 2012, Full version: http://control.ee.ethz.ch/index.cgi?page=publications;action=details;id=4196, pp. 3433–3438.

[MVM$^+$10] Peyman Mohajerin Esfahani, Maria Vrakopoulou, Kostas Margellos, John Lygeros, and Goran Andersson, *Cyber attack in a two-area power system: Impact identification using reachability*, American Control Conference, 2010, pp. 962–967.

[MVM$^+$11] _____, *A robust policy for automatic generation control cyber attack in two area power network*, 49th IEEE Conference Decision and Control, 2011, pp. 5973–5978.

[MVW89]   Mohammad-Ali Massoumnia, G. C. Verghese, and A. S. Willsky, *Failure detection and identification*, IEEE Transaction on Automatic Control **34** (1989), no. 3, 316–321.

[NF06]   Mattias Nyberg and Erik Frisk, *Residual generation for fault diagnosis of system described by linear differential-algebraic equations*, IEEE Transaction on Automatic Control **51** (2006), no. 12, 1995–2000.

[PI01]   Claudio De Persis and Alberto Isidori, *A geometric approach to nonlinear fault detection and isolation*, IEEE Trans. Automat. Control **46** (2001), no. 6, 853–865.

[PW98]   Jan Willem Polderman and Jan C. Willems, *Introduction to mathematical systems theory*, Texts in Applied Mathematics, vol. 26, Springer-Verlag, New York, 1998, A behavioral approach.

[ref]   *Power systems test case archive, college of engineering,university of washington*, URL: http://www.ee.washington.edu/research/pstca/.

[SF91]   R. Seliger and P.M. Frank, *Fault diagnosis by disturbance-decoupled nonlinear observers*, Proceedings of the 30th IEEE Conference on Decision and Control, 1991, pp. 2248–2253.

[Shc07]   A. A. Shcheglova, *Nonlinear differential-algebraic systems*, Sibirsk. Mat. Zh. **48** (2007), no. 4, 931–948. MR 2355385 (2009c:34002)

[SMEKL13] Bratislav Svetozarevic, Peyman Mohajerin Esfahani, Maryam Kamgarpour, and John Lygeros, *A robust fault detection and isolation filter for a horizontal axis variable speed wind turbine*, American Control Conference (ACC), 2013, June 2013, pp. 4453–4458.

[ZD97]   Kemin Zhou and John C. Doyle, *Essentials of robust control*, Prentice Hall, September 1997.