2024 European Control Conference (ECC)
June 25-28, 2024. Stockholm, Sweden

Sensor Fault Detection and Isolation in Autonomous Nonlinear Systems Using Neural Network-Based Observers

John Cao* M. Umar B. Niazi† Matthieu Barreau* Karl Henrik Johansson*

***Abstract*—This paper presents a novel observer-based approach to detect and isolate faulty sensors in nonlinear systems. The proposed sensor fault detection and isolation (s-FDI) method applies to a general class of nonlinear systems. Our focus is on s-FDI for two types of faults: complete failure and sensor degradation. The key aspect of this approach lies in the utilization of a neural network-based Kazantzis-Kravaris/Luenberger (KKL) observer. The neural network is trained to learn the dynamics of the observer, enabling accurate output predictions of the system. Sensor faults are detected by comparing the actual output measurements with the predicted values. If the difference surpasses a theoretical threshold, a sensor fault is detected. To identify and isolate which sensor is faulty, we compare the numerical difference of each sensor measurement with an empirically derived threshold. We derive both theoretical and empirical thresholds for detection and isolation, respectively. Notably, the proposed approach is robust to measurement noise and system uncertainties. Its effectiveness is demonstrated through numerical simulations of sensor faults in a network of Kuramoto oscillators.**

I. Introduction

Sensor fault detection and isolation (s-FDI) plays a pivotal role in ensuring the safe and efficient operation of numerous industrial processes. We address two distinct types of sensor faults: complete failure and sensor degradation. Complete failure occurs when a sensor becomes entirely nonfunctional, often due to mechanical breakdown or similar. In contrast, sensor degradation results in a gradual decline in the sensor's measurement accuracy. When left undetected, these sensor faults can lead to disruptive consequences for the system. Effective s-FDI methods serve as a proactive solution, allowing system operators to detect sensor faults early, localize the faulty sensor, and take corrective action before they escalate into issues that could result in costly damage or downtime.

The most widely adapted s-FDI methods to date are primarily based on the concept of *analytical redundancy*, which use the principle of residual generation to compare the variance between a predicted system output to the actual measurement. Historically, these methods rely on explicit mathematical models of the system under consideration, as noted in [1]. This model-based approach was initially developed for linear systems in the 1970s, exemplified by pioneering work such as [2], which demonstrated the feasibility of designing filters for detecting and localizing faults within observable system dynamics. Subsequent refinements and enhancements, as seen in [3], resulted in the famous "Beard-Jones Fault Detection filter."

Parallel to these developments, the framework of observer-based fault detection schemes emerged for linear systems, with an early reference being [4]. Over time, this approach has gained recognition as one of the most successful methods for s-FDI, leading to diverse research directions. For instance, the application of sliding-mode observers for s-FDI by [5]–[7] allowed explicit reconstruction of the sensor faults by manipulating the output injection error. Nonlinear unknown input observers have also been prevalent for s-FDI [1]. In recent years, interval-based unknown input observers have gained prominence. These observers rely on relaxed assumptions on system inputs and nonlinearities [8]–[11].

Another principal approach to analytical redundancy-based s-FDI is by using data-driven methods. These methods do not require explicit system models, relying instead on sensor data to approximate the underlying dynamics and generate residuals, as highlighted in [12]. For instance, [13] demonstrated s-FDI using long short-term memory neural networks (LSTM), where residuals were formed by comparing network predictions based on past time-series data with actual measurements. Similar neural network-based approaches have been proposed for various domains, including industrial manufacturing processes, power plants, and unmanned aerial vehicles, as indicated in [12], [14], [15].

Despite the rich diversity of existing s-FDI methods, it's important to acknowledge that, to the best of the authors' knowledge, these techniques are often challenging to implement in practice. Observer-based methods are constrained by their assumptions about specific system structures, while data-driven approaches require substantial sensor data, which can be both difficult and costly to obtain.

In this paper, we tackle these challenges by introducing a novel learning-based approach to s-FDI using neural network-based nonlinear Luenberger observers, specifically Kazantzis-Kravaris/Luenberger (KKL) observers [16],

* Division of Decision and Control Systems, Digital Futures, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden. Emails: johncao@kth.se, barreau@kth.se, kallej@kth.se

† Laboratory for Information & Decision Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Email: niazi@mit.edu

This work is supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation, Sweden. It has also received funding from the European Union's Horizon Research and Innovation Programme under Marie Skłodowska-Curie grant agreement No. 101062523.

[17]. KKL observers lift the original system to a higher-dimensional state space, where the system behaves linearly and is stable. The nonlinear transformation required by the lift is obtained by solving a certain partial differential equation. The observer's design is based on this transformed system. To estimate the system's state, an inverse transformation is applied, to obtain the estimate in the original state space. A notable advantage of KKL observers is their flexibility; they do not rely on specific triangular structures or normal forms of the system but rather depend on relatively mild observability conditions, making them applicable to a broad class of nonlinear systems.

Our main contributions in this paper are

1) We develop a novel s-FDI algorithm using a neural network-based KKL observer with no assumptions on the system's structure.
2) We derive a theoretical threshold for sensor fault detection based on the residual, and devise an empirical method to obtain a threshold for sensor fault isolation.
3) Through simulations, we show that the proposed method can effectively detect and isolate sensor faults under various circumstances while remaining robust to model uncertainties and measurement noise.

The outline of this paper is as follows. Section II formulates the s-FDI problem under some mild assumptions. Section III presents the s-FDI method. Numerical results are provided in Section IV with various fault cases. Lastly, Section V concludes the paper.

II. Problem Formulation

We consider a nonlinear system

$$\dot{x}(t) = f(x(t)) + w(t) \tag{1a}$$
$$y(t) = \phi(t)\left[h(x(t)) + v(t) + \zeta(t)\right] \tag{1b}$$

where $x(t) \in \mathcal{X} \subset \mathbb{R}^{n_x}$ is the state, $y(t) \in \mathbb{R}^{n_y}$ is the output, $f : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ and $h : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y}$ are smooth maps, and $w(t), v(t)$ are process and measurement noises, respectively. The measured output (1b) might be affected by sensor faults $\phi(t)$ and $\zeta(t)$, where

$$\phi(t) = \begin{bmatrix}\phi_1(t) & \dots & \phi_{n_y}(t)\end{bmatrix}^{\mathrm{T}} \in \{0,1\}^{n_y}$$

models the complete failure of sensor i when $\phi_i(t) = 0$, and

$$\zeta(t) = \begin{bmatrix}\zeta_1(t) & \dots & \zeta_{n_y}(t)\end{bmatrix}^{\mathrm{T}} \in \mathbb{R}^{n_y}$$

models the degradation of sensor i, affecting its measurement accuracy, when $\zeta(t) \neq 0$.

Note that a system with fault-free sensors at time t will have $\phi(t) = 1_{n_y}$ and $\zeta(t) = 0_{n_y}$. Any other value of these signals represents a type of fault, such as sensor bias or degradation when $\zeta_i(t) \neq 0$, for $t \geq T$, or sensor failure when $\phi_i(t) = 0$, for $t \geq T$, where $i \in \{1, \dots, n_y\}$ and $T \in \mathbb{R}_{>0}$ is a time at which fault occurs.

Assumption 1. We assume the following regarding the process and measurement noise:

(i) Essential boundedness: $\|w\|_{L^\infty} \leq \bar{w}$ and $\|v\|_{L^\infty} \leq \bar{v}$, where $\bar{w}, \bar{v} > 0$ are known and $\|\cdot\|_{L^\infty}$ denotes the essential supremum norm.

(ii) Bounded effect:

$$\exists \psi \in \mathcal{K}_\infty, \quad \|x(t; x_0, w) - x(t; x_0, 0)\| \leq \psi(\bar{w})$$

where $x(t; x_0, w)$ denotes the state trajectory of (1a) initialized at $x(0) = x_0$ and driven by the noise $w(t)$.

Remark 1. Assumption 1(i) is a standard assumption in the robust state estimation literature. It says that the signals $w(t)$ and $v(t)$ almost always remain bounded, and the instances at which $w(t)$ or $v(t)$ go unbounded are of zero Lebesgue measure. On the other hand, Assumption 1(ii) requires that the model f without a process noise does a good job of describing the noisy system (1a). Such guarantees are usually provided in the system identification literature, and the reader is referred to [18], [19] for more details. ◇

Given a system (1), a neural network-based Kazantzis-Kravaris/Luenberger (NN-KKL) observer [20] is given by

$$\dot{\hat{z}}(t) = A\hat{z}(t) + By(t) \tag{2a}$$
$$\hat{x}(t) = \hat{\mathcal{T}}^*_\eta(\hat{z}(t)) \tag{2b}$$
$$\hat{y}(t) = h(\hat{x}(t)) \tag{2c}$$

which takes output measurements $y(t)$ as input and gives the estimated state $\hat{x}(t)$ and predicted output $\hat{y}(t)$. In (2a), $A \in \mathbb{R}^{n_z \times n_z}$ is a Hurwitz matrix and $B \in \mathbb{R}^{n_z \times n_y}$ is such that the pair (A, B) is controllable. The observer state $\hat{z}(t) \in \mathbb{R}^{n_z}$ follows a nonlinear transformation $\hat{z} = \hat{\mathcal{T}}_\theta(\hat{x})$. The neural network $\hat{\mathcal{T}}_\theta$ with parameters θ is an approximation of the injective map $\mathcal{T} : \mathcal{X} \to \mathbb{R}^{n_y(2n_x+1)}$. On the other hand, in (2b), $\hat{\mathcal{T}}^*_\eta$ is a neural network approximation with parameters η of $\mathcal{T}^* : \mathbb{R}^{n_z} \to \mathcal{X}$, which is the left inverse of $\mathcal{T}$. Note that the observer is trained by considering fault-free sensors. We refer to [20] for details on the design of NN-KKL observers.

Assumption 2. We assume the following regarding nonlinear transformations $\mathcal{T}, \mathcal{T}^*$ and their corresponding neural network approximations $\hat{\mathcal{T}}_\theta, \hat{\mathcal{T}}^*_\eta$:

(i) Lipschitzness: $\hat{\mathcal{T}}_\theta$ and $\hat{\mathcal{T}}^*_\eta$ are Lipschitz continuous over $\mathcal{X}$ and $\mathcal{Z}$, respectively, where $\mathcal{Z} \supseteq \mathcal{T}(\mathcal{X})$, i.e.

$$\exists \ell_\theta \in \mathbb{R}_{>0},\ \forall x, \hat{x} \in \mathcal{X}, \quad \|\hat{\mathcal{T}}_\theta(x) - \hat{\mathcal{T}}_\theta(\hat{x})\| \leq \ell_\theta \|x - \hat{x}\|,$$
$$\exists \ell_\eta \in \mathbb{R}_{>0},\ \forall z, \hat{z} \in \mathcal{Z}, \quad \|\hat{\mathcal{T}}^*_\eta(z) - \hat{\mathcal{T}}^*_\eta(\hat{z})\| \leq \ell_\eta \|z - \hat{z}\|.$$

(ii) Uniform injectivity:

$$\exists \rho \in \mathcal{K},\ \forall x, \hat{x} \in \mathcal{X}, \quad \|x - \hat{x}\| \leq \rho(\|\mathcal{T}(x) - \mathcal{T}(\hat{x})\|),$$
$$\exists \rho^* \in \mathcal{K}, \forall z, \hat{z} \in \mathcal{Z}, \quad \|z - \hat{z}\| \leq \rho^*(\|\mathcal{T}^*(z) - \mathcal{T}^*(\hat{z})\|).$$

Remark 2. Assumption 2(i) is satisfied if the activation function of neural networks is chosen to be Lipschitz continuous. Assumption 2(ii) relates to the existence of a KKL observer [17]. Notice that uniformly injective $\mathcal{T}$ implies the uniform injectivity of its inverse $\mathcal{T}^*$. It is important to remark that Assumption 2(i) and (ii) imply boundedness of approximation errors, i.e.,

$$\sup_{x\in\mathcal{X}} \|\mathcal{T}(x) - \hat{\mathcal{T}}_\theta(x)\| < \infty, \quad \sup_{z\in\mathcal{Z}} \|\mathcal{T}^*(z) - \hat{\mathcal{T}}^*_\eta(z)\| < \infty$$

where $\mathbb{R}^{n_z} \supset \mathcal{Z} \supseteq \mathcal{T}(\mathcal{X})$. This holds because

$$\|\mathcal{T}(x) - \hat{\mathcal{T}}_\theta(x)\| \le \|\mathcal{T}(x)\| + \|\hat{\mathcal{T}}_\theta(x)\| \le \rho^*(\|x\|) + \ell_\theta\|x\|$$

and $\mathcal{X} \subset \mathbb{R}^{n_x}$ is a compact set. Further improvement on this bound can be achieved using the universal approximation property of neural networks given enough data is generated and an appropriate network architecture is chosen. ⋄

Observer-based s-FDI of nonlinear systems relies on the design of observers that are accurate in state estimation and output prediction. However, for general nonlinear systems, designing observers is a challenging task. Therefore, we consider an NN-KKL observer (2) proposed in [20], which is designed using a physics-informed learning approach. The focus of the present paper is on developing an s-FDI method for general nonlinear systems using an NN-KKL observer (2). Specifically, we address the following problems:

(P1) Detection: How to detect whenever a fault occurs in one or more sensors?

(P2) Isolation: How to identify which sensor is faulty?

To solve the problems stated above, we define residuals to compare a measured output from the system with a predicted output from the NN-KKL observer. The residual corresponding to ith sensor is defined as

$$\begin{aligned} r_i(t) &\doteq |y_i(t) - \hat{y}_i(t)| \\ &= |\phi_i(t)[h_i(x(t)) + v_i(t) + \zeta_i(t)] - h_i(\hat{x}(t))|. \end{aligned} \tag{3}$$

Consider equally-distant discrete time samples $t_0, t_1, t_2, \ldots$ with $\delta = t_k - t_{k-1}$ for $k \in \mathbb{N}$, then we define a differentiated residual of ith sensor as the absolute value of numerical differentiation of the residual, i.e.,

$$\tilde{r}_i(t_k) \doteq \frac{1}{\delta}\left|r_i(t_k) - r_i(t_{k-1})\right|. \tag{4}$$

Let $\tau_i \in \mathbb{R}_{>0}$ denote a threshold for residual i such that, in the steady state, $r_i(t) \le \tau_i$ when there are no faults. Let $r_\Delta \in \mathbb{R}_{>0}$ denote a threshold for $\tilde{r}_i$ such that, in the steady state, $\tilde{r}_i(t) \le r_\Delta$ when sensor i is not faulty.

III. s-FDI using NN-KKL Observers

In this section, we first derive theoretical bounds for the sensor residuals. Then, we devise an empirical method to compute the threshold for differentiated residuals. Finally, we present the proposed s-FDI method.

A. Upper bounds of the residuals

We derive two upper bounds on the residuals[1]. The first bound is straightforward but requires an additional stricter assumption that both $\hat{\mathcal{T}}_\theta$ and $\hat{\mathcal{T}}^*_\eta$ are contractions, i.e., their Lipschitz constants are less than 1. The second bound is more general and relies on Assumption 1 and 2.

[1]The proofs are available in the extended version.

Proposition 1. *Suppose the neural networks $\hat{\mathcal{T}}_\theta$ and $\hat{\mathcal{T}}^*_\eta$ are Lipschitz continuous with $\ell_\theta \in [0,1)$ and $\ell_\eta \in [0,1)$. Then, in a fault-less case,*

$$r_i(t) \le \ell_{h_i}\frac{\ell_\eta\xi(x) + \xi^*(z)}{1 - \ell_\theta\ell_\eta} + \bar{v} \tag{5}$$

where ℓ_{h_i} is the Lipschitz constant of $h_i(x)$ for $x \in \mathcal{X}$, $\bar{v}$ is given in Assumption 1(i), and

$$\xi(x) \doteq \mathcal{T}(x) - \hat{\mathcal{T}}_\theta(x), \text{ and } \xi^*(z) \doteq \mathcal{T}^*(z) - \hat{\mathcal{T}}^*_\eta(z) \tag{6}$$

which are bounded for $x \in \mathcal{X}$ and $z \in \mathcal{Z}$ by Assumption 2(ii).

Remark 3. The assumption of Proposition 1 that both Lipschitz constants are less than one can be satisfied by regularizing the weights of neural networks during training. For instance, an l-layer ReLU network with $W^{(1)}, \ldots, W^{(l)}$, which are the weight matrices of the neural network layers, has a Lipschitz constant $\ell = \|W^{(1)}\|\ldots\|W^{(l)}\|$. By regularizing weights such that the maximum singular value of each weight matrix is less than one, it can be ensured that $\ell < 1$. ⋄

Recall $z = \mathcal{T}(x)$ and $\hat{z} = \hat{\mathcal{T}}_\theta(x)$, where

$$\dot{z}(t) = Az(t) + Bh(\bar{x}(t))$$

with $\bar{x}(t) \doteq x(t; x_0, 0)$ the noise-free state trajectory, and the dynamics of $\hat{z}(t)$ is given in (2a).

Lemma 1. *Suppose the matrix A is Hurwitz and diagonalizable with eigenvalue decomposition $A = V\Lambda V^{-1}$. Then, the error $\tilde{z}(t) := z(t) - \hat{z}(t)$ satisfies*

$$\|\tilde{z}(t)\| \le \|\tilde{z}_0\|\kappa(V)e^{-ct}$$

where $c > 0$, ψ in Assumption 1(ii), and ℓ_h is the Lipschitz constant of $h(x)$ for $x \in \mathcal{X}$.

Using the above lemmas, we can obtain a general upper bound on the residuals under Assumption 1 and 2.

Proposition 2. *Suppose the matrix A is diagonalizable with eigenvalue decomposition $A = V\Lambda V^{-1}$. Then, in a fault-less case,*

$$\begin{aligned} r_i(t) \le \bar{v} + \ell_{h_i}\Big[&\xi^*(z) + \ell_\eta\kappa(V)e^{-ct}\|\tilde{z}_0\| \\ &+ \tfrac{\kappa(V)}{c}\|B\|(1 - e^{-ct})(\ell_h\psi(\bar{w}) + \sqrt{n_y}\bar{v})\Big] \end{aligned} \tag{7}$$

where $\bar{v}$ is from Assumption 1(i), ℓ_{h_i} is the Lipschitz constant of $h_i(x)$ for $x \in \mathcal{X}$, $\xi^*(z)$ is defined in (6), $\kappa(V)$ is the condition number of V, and $\tilde{z}_0 = \tilde{z}(0)$.

B. Fault detection by computing a theoretical threshold for the residuals

After learning the NN-KKL observer on the training datasets $\mathcal{X}_{\text{train}} \subset \mathcal{X}$ and $\mathcal{Z}_{\text{train}} \subset \mathcal{Z}$, generate a test dataset $\mathcal{X}_{\text{test}} \subset \mathcal{X}$ by simulating a noise-less ($w \equiv 0, v \equiv 0$) and fault-free ($\phi \equiv 1, \zeta \equiv 0$) system (1a). Then, following the methodology presented in [20], we can generate $\mathcal{Z}_{\text{test}} \approx \mathcal{T}(\mathcal{X}_{\text{test}})$, where $\mathcal{Z}_{\text{test}} \subset \mathcal{Z}$. The datasets $\mathcal{X}_{\text{test}}$ and $\mathcal{Z}_{\text{test}}$ are then used to estimate the approximation errors incurred by $\hat{\mathcal{T}}_\theta$ and $\hat{\mathcal{T}}_\eta^*$. Define

$$\hat{\epsilon} \doteq \max_{\substack{x\in\mathcal{X}_{\text{test}},\\ z\in\mathcal{Z}_{\text{test}}}} \left\| z - \hat{\mathcal{T}}_\theta(x) \right\|, \quad \hat{\epsilon}^* \doteq \max_{\substack{x\in\mathcal{X}_{\text{test}},\\ z\in\mathcal{Z}_{\text{test}}}} \left\| x - \hat{\mathcal{T}}_\eta^*(z) \right\|.$$

Then, Proposition 2 leads to

$$\limsup_{t\to\infty} r_i(t) \leq \bar{v} + \ell_{h_i}\left[\xi^*(z) + \frac{\kappa(V)}{c}\|B\|(\ell_h\psi(\bar{w}) + \sqrt{n_y}\bar{v})\right].$$

Therefore, in this case, we can choose the threshold to be

$$\boxed{\tau_i = \bar{v} + \ell_{h_i}\left[\hat{\epsilon}^* + \frac{\kappa(V)}{c}\|B\|(\ell_h\psi(\bar{w}) + \sqrt{n_y}\bar{v})\right].} \quad (8)$$

Fault detection: At a steady state when t becomes large, it holds that $r_i(t) \leq \tau_i$ when there are no sensor faults. When the residual $r_i(t)$ surpasses τ_i, it signifies that the measured output is significantly different from the predicted output. Since τ_i has been computed using tight inequalities, sensor faults can be detected by measuring the residuals $r_i(t)$ and raising an alert whenever $r_i(t) > \tau_i$ for any i and any $t \geq 0$.

C. Fault isolation by computing an empirical threshold for differentiated residuals

As described in the previous subsection, a sensor fault is detected whenever the residual $r_i(t)$ surpasses the threshold τ_i. However, $r_i(t)$ surpassing τ_i and having the largest value among other residuals do not mean that sensor i is faulty. This is because a fault in sensor j distorts the output $y_j(t)$, which is filtered through the observer dynamics and then transformed by $(h_j \circ \hat{\mathcal{T}}_\eta^*)(\cdot)$. Since the neural network $\hat{\mathcal{T}}_\eta^*$ is not diagonal, a fault in sensor j could induce a large residual $r_i(t)$ even when sensor i is not faulty. This phenomenon can also be observed in the third row of Fig. 2 in Section IV. Because of the inter-dependencies due to non-diagonal $\hat{\mathcal{T}}_\eta^*$, the transients after the occurrence of a fault may persist above the threshold, leading to an inability to exactly isolate the sensor faults from the residuals. Therefore, fault isolation is done by evaluating differentiated residuals $\tilde{r}_i(t_k)$.

Based on our empirical observations, a spike is generated in the differentiated residual $\tilde{r}_i(t_k)$ whenever an abrupt fault ($\phi_i(t) = 0$ or $\zeta_i(t) \neq 0$) is introduced in sensor i at time t_k. The explanation for this spike is that the fault in the output gets numerically differentiated. However, isolation is tedious when the fault signal $\zeta_i(t)$ is very smooth and of low magnitude. Nevertheless, it is reasonable to assume that faults are irregular and non-smooth signals.

Consider the test datasets $\mathcal{X}_{\text{test}}$ and $\mathcal{Z}_{\text{test}}$, where we denote the state trajectories $x^j(t_k) \doteq x(t_k; x_0^j, 0)$ that are initialized at p initial points $x(0) = x_0^j$, for $j = 1, \ldots, p$, and the corresponding observer estimates as $\hat{x}^j(t_k) \doteq \hat{x}(t_k; \hat{\mathcal{T}}_\eta^*(z_0^j))$. We simulate the NN-KKL observer and compute the residuals $r_1^j(t_k), \ldots, r_{n_y}^j(t_k)$, for $k = 0, \ldots, T$, where $T \in \mathbb{N}$ and

$$r_i^j(t_k) = |y_i(t_k; x^j) - \hat{y}_i(t_k; \hat{x}^j)|.$$

Here, $t_T > 0$ is chosen large enough to allow the observer to converge to a neighborhood of the original state. Then, the corresponding differentiated residuals $\tilde{r}_1^j(t_k), \ldots, \tilde{r}_{n_y}^j(t_k)$, for $k = 1, \ldots, T$, are obtained using (4). Let t_c, for $c < T$, denote the minimum time at which the observer is in the steady state. Then an empirical threshold for differentiated residuals is computed as follows:

$$\boxed{r_\Delta = \max_{\substack{k\in\{c,\ldots,T\},\ j\in\{1,\ldots,p\}\\ i\in\{1,\ldots,n_y\}}} \tilde{r}_i^j(t_k).} \quad (9)$$

Fault isolation: Given that the learned NN-KKL observer estimates the system state accurately, it holds that $\tilde{r}_i(t) \leq r_\Delta$ at steady state when t_k is large enough. Therefore, when the differentiated residual $\tilde{r}_i(t)$ surpasses the empirical threshold r_Δ, an alert can be raised that sensor i is probably faulty.

IV. Simulation Results

Our approach is demonstrated in this section using numerical simulations. We show that s-FDI of a highly nonlinear system can be achieved using NN-KKL observers, even in the presence of additive process and measurement noises. Both types of faults, sensor degradation represented by the fault signal $\zeta_i(t) \neq 0$ and sensor failure represented by $\phi_i(t) = 0$ in the system output (1b), are demonstrated.

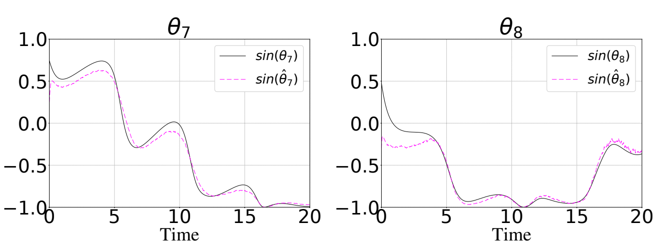


Fig. 1: Estimated and true trajectories of states θ_7 and θ_8 under the influence of process and sensor noise.

A. Kuramoto Model

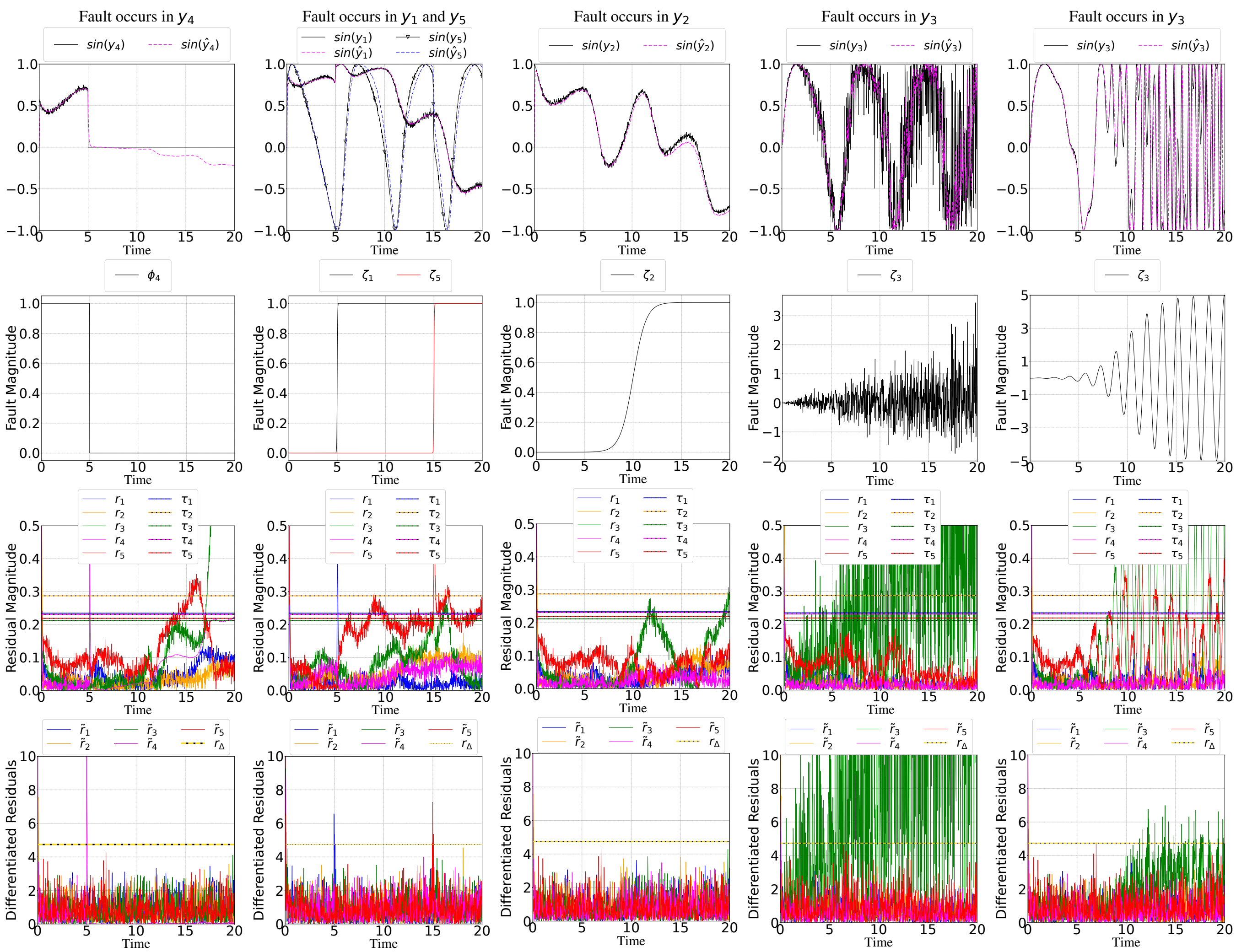
We consider the Kuramoto model for demonstrating our s-FDI method. The Kuramoto model describes the phenomena of synchronization in a multitude of systems, including electric power networks, multi-agent coordination, and distributed software networks [21]. The dynamics of a network with n nodes are given as

$$\dot{\theta}_i(t) = \omega_i + \textstyle\sum_{j=1}^n a_{ij}\sin(\theta_i(t) - \theta_j(t)) \quad (10)$$

where $\theta_i(t) \in \mathbb{S}^1$ is the phase angle of node $i = 1, .., n$, $\omega_i \in \mathbb{R}$ is the natural frequency of node i, and $a_{ij} \geq 0$ denotes the coupling between node i and j. The state trajectories of (10) are often represented graphically as $\sin(\theta_i)$ to better illustrate their synchronization.

B. Experimental Setup

For (10), we consider a network of 10 nodes with randomly generated natural frequencies ω_i and couplings a_{ij}. The measurements are chosen as $y = \begin{bmatrix}\theta_1 & \theta_2 & \theta_3 & \theta_4 & \theta_5\end{bmatrix}^\intercal$. A set of 50 initial conditions is generated randomly in $\mathcal{X}_{\text{train}} \subset$



(a) (b) (c) (d) (e)

Fig. 2: (a) Complete failure of sensor 4. (b) Degradation of sensors 1 and 5 at different time points with abrupt faults. (c) Smooth sigmoidal fault in sensor 2. (d) Gradually increasing white noise in sensor 3. (e) Gradually increasing sinusoidal signal in sensor 3.

$[-2,2]^{10}$. We choose Runge-Kutta-4 as our numerical ODE solver to simulate (10) and (2a) over a time interval of $[0,30]$, partitioned into 4000 sample points for each trajectory. The neural network $\hat{\mathcal{T}}^*_\eta$ is chosen as a dense feed-forward network, consisting of 3 hidden layers of 250 neurons with ReLU activation function. Model training is facilitated by data standardization and learning rate scheduling. Following [22], the matrices of (2a) are chosen as

$$A = \Lambda \otimes I_{n_y}, \qquad B = \Gamma \otimes I_{n_y}$$

where $\Lambda \in \mathbb{R}^{(2n_x+1)\times(2n_x+1)}$ is a diagonal matrix with diagonal elements linearly distributed in $[-15,-21]$, $\Gamma \in \mathbb{R}^{2n_x+1}$ is a column vector of ones, and I_{n_y} is the identity matrix of size $n_y \times n_y$. Here, n_x and n_y are 10 and 5, respectively, and $n_z = n_y(2n_x+1) = 105$.

The estimation capabilities of the observer under fault-free conditions are demonstrated in Fig. 1, which shows the estimated and true trajectories of two (randomly chosen) unmeasured states θ_7 and θ_8 over a time interval of $[0,20]$, with noise terms $w(t), v(t) \sim \mathcal{N}(0, 0.02)$. The figure demonstrates that the estimation error is stable under noise and neural network approximation error.

C. Numerical Results

We now apply the learned neural network-based KKL observer to perform s-FDI. The theoretical thresholds τ_i of the residuals $r_i(t)$ computed by (8) are shown in the third row of Fig. 2. On the other hand, the empirical threshold $r_\Delta = 4.74$ is computed using (9) according to the method described in Section III-C. We choose $N = 100$ initial conditions to create $\mathcal{X}_{\text{test}} \in [-2,2]^{10}$ and to compute r_Δ using (9). Fig. 2c–2e demonstrate the detection and isolation capabilities of our method under various faults. In the figure, the rows correspond to the measured and estimated state trajectories, the fault signal, the residuals $r_i = |y_i - \hat{y}_i|$ and the finite difference approximation (4) respectively.

In Fig. 2a, we show that our method is capable of detecting sensor shutdowns due to complete failure, which we demonstrate by modeling the fault in sensor 4 with $\phi_4(t) = 0$. Fig. 2b illustrates the situation when more than one fault is present. Sensors 1 and 5 are disturbed by $\zeta_1(t) = 1$, at $t \geq 5$, and $\zeta_5(t) = 1$, at $t \geq 15$. Each fault is distinctly detectable at the moment of occurrence.

In Fig. 2c, sensor 2 is disturbed by a smooth sigmoidal fault term $\zeta_2(t)$ introduced at $t = 5$. Because of the smooth sigmoidal signal as a fault, the observer does not detect any anomaly and continues to follow the measured state trajectories, thus generating a small residual. Although introducing an abrupt fault as in Fig. 2a and 2b induces a transient in the observer, causing a large residual to be generated, introducing a smooth fault in Fig. 2c resulted in only detection but not isolation. Due to the stability of the observer, it attempts to track the faulty trajectories after the occurrence of the fault, leading to a decrease in the residual magnitudes subsequently.

Fig. 2d and 2e illustrate the case when the fault signal on sensor 3 is introduced gradually. In Fig. 2d, we simulate a fault in sensor y_3 which is a gradually increasing white noise. In Fig. 2e, the fault is a sinusoidal signal that gradually increases its magnitude from 0 to 5. Here, due to a gradual increase in the fault signals' magnitudes, it can be observed that both detection and isolation are successfully performed but with a small delay.

V. Conclusion and Future Work

We have introduced a novel sensor fault detection and isolation method tailored for a general class of nonlinear systems. For s-FDI, we leverage a neural network-based Kazantzis-Kravaris/Luenberger (KKL) observer for residual generation. We derived theoretical upper bounds for the residuals obtained by comparing measured outputs with those predicted by the observer. These upper bounds serve as the foundation for analytically computing thresholds. When a residual crosses its corresponding threshold, it indicates the detection of sensor faults. However, the critical task of fault isolation relies on the numerically differentiated residuals rather than the usual residuals. To this end, we introduced an empirical methodology, involving experimentation with the learned KKL observer, to calculate the threshold for the differentiated residuals.

We demonstrated the efficacy of our approach on a network of Kuramoto oscillators by evaluating various fault scenarios, including sensor degradation and complete failure. The comprehensive set of simulations provides compelling evidence of the method's robustness and effectiveness in fault detection and isolation within autonomous nonlinear systems.

The theory of KKL observers extends to non-autonomous systems, and adapting our method to those systems remains an open research topic. It is also of interest to study the performance of the method in real applications, especially in systems where conventional solutions are known to fail.

References

[1] M. Thirumarimurugan, N. Bagyalakshmi, and P. Paarkavi, "Comparison of fault detection and isolation methods: A review," in *2016 10th International Conference on Intelligent Systems and Control (ISCO)*, 2016, pp. 1–6.

[2] R. V. Beard, "Failure accomodation in linear systems through self-reorganization," Ph.D. dissertation, Massachusetts Institute of Technology, 1971.

[3] H. L. Jones, "Failed detection in linear systems," Ph.D. dissertation, Massachusetts Institute of Technology, 1973.

[4] R. N. Clark, D. C. Fosth, and V. M. Walton, "Detecting instrument malfunctions in control systems," *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-11, no. 4, pp. 465–473, 1975.

[5] C. Edwards, S. K. Spurgeon, and R. J. Patton, "Sliding mode observers for fault detection and isolation," *Automatica*, vol. 36, no. 4, pp. 541–553, 2000.

[6] X.-G. Yan and C. Edwards, "Nonlinear robust fault reconstruction and estimation using a sliding mode observer," *Automatica*, vol. 43, no. 9, pp. 1605–1614, 2007.

[7] C. P. Tan and C. Edwards, "Sliding mode observers for robust detection and reconstruction of actuator and sensor faults," *International Journal of Robust and Nonlinear Control*, vol. 13, no. 5, pp. 443–463, 2003.

[8] F. Zhu, Y. Tang, and Z. Wang, "Interval-observer-based fault detection and isolation design for t-s fuzzy system based on zonotope analysis," *IEEE Transactions on Fuzzy Systems*, vol. 30, no. 4, pp. 945–955, 2022.

[9] F. Xu, J. Tan, X. Wang, and B. Liang, "Conservatism comparison of set-based robust fault detection methods: Set-theoretic UIO and interval observer cases," *Automatica*, vol. 105, pp. 307–313, 2019.

[10] Z.-H. Zhang and G.-H. Yang, "Event-triggered fault detection for a class of discrete-time linear systems using interval observers," *ISA Transactions*, vol. 68, pp. 160–169, 2017.

[11] Q. Su, Z. Fan, T. Lu, Y. Long, and J. Li, "Fault detection for switched systems with all modes unstable based on interval observer," *Information Sciences*, vol. 517, pp. 167–182, 2020.

[12] R. Iqbal, T. Maniak, F. Doctor, and C. Karyotis, "Fault detection and isolation in industrial processes using deep learning approaches," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 5, pp. 3077–3084, 2019.

[13] J. Yang, Y. Guo, and W. Zhao, "Long short-term memory neural network based fault detection and isolation for electro-mechanical actuators," *Neurocomputing*, vol. 360, pp. 85–96, 2019.

[14] H. S. Farahani, A. Fatehi, and M. A. Shoorehdeli, "On the application of domain adversarial neural network to fault detection and isolation in power plants," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2020, pp. 1132–1138.

[15] D. Guo, M. Zhong, H. Ji, Y. Liu, and R. Yang, "A hybrid feature model and deep learning based fault diagnosis for unmanned aerial vehicle sensors," *Neurocomputing*, vol. 319, pp. 155–163, 2018.

[16] N. Kazantzis and C. Kravaris, "Nonlinear observer design using lyapunov's auxiliary theorem," *Systems & Control Letters*, vol. 34, no. 5, pp. 241–247, 1998.

[17] V. Andrieu and L. Praly, "On the existence of a Kazantzis–Kravaris/Luenberger observer," *SIAM Journal on Control and Optimization*, vol. 45, no. 2, pp. 432–456, 2006.

[18] M. Milanese, J. Norton, H. Piet-Lahanier, and É. Walter, *Bounding Approaches to System Identification*. Springer New York, NY, 1996.

[19] H. Mania, M. I. Jordan, and B. Recht, "Active learning for nonlinear system identification with guarantees." *Journal of Machine Learning Research*, vol. 23, pp. 1–30, 2022.

[20] M. U. B. Niazi, J. Cao, X. Sun, A. Das, and K. H. Johansson, "Learning-based design of luenberger observers for autonomous nonlinear systems," in *2023 American Control Conference (ACC)*, 2023, pp. 3048–3055.

[21] F. Dörfler and F. Bullo, "Synchronization in complex networks of phase oscillators: A survey," *Automatica*, vol. 50, no. 6, pp. 1539–1564, 2014.

[22] P. Bernard, V. Andrieu, and D. Astolfi, "Observer design for continuous-time dynamical systems," *Annual Reviews in Control*, vol. 53, pp. 224–248, 2022.