International JOURNAL of CONTROL

Taylor & Francis
Taylor & Francis Group

International Journal of Control

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tcon20

Observer-based safety monitoring of nonlinear dynamical systems with neural networks via quadratic constraint approach

Tao Wang, Yapeng Li, Zihao Mo, Wesley Cooke & Weiming Xiang

INTERNATIONAL JOURNAL OF CONTROL
2024, VOL. 97, NO. 10, 2420–2432
https://doi.org/10.1080/00207179.2023.2274924

Taylor & Francis
Taylor & Francis Group

Observer-based safety monitoring of nonlinear dynamical systems with neural networks via quadratic constraint approach

Tao Wang[a], Yapeng Li[a], Zihao Mo[b], Wesley Cooke[b] and Weiming Xiang[b]

[a]School of Electrical Engineering, Southwest Jiaotong University, Chengdu, People's Republic of China; [b]School of Computer and Cyber Sciences, Augusta University, Augusta, GA, USA

ABSTRACT
The safety monitoring for nonlinear dynamical systems with embedded neural network components is addressed in this paper. The interval-observer-based safety monitor is developed consisting of two auxiliary neural networks derived from the neural network components of the dynamical system. Due to the presence of nonlinear activation functions in neural networks, we use quadratic constraints on the global sector to abstract the nonlinear activation functions in neural networks. By combining a quadratic constraint approach for the activation function with Lyapunov theory, the interval observer design problem is transformed into a series of quadratic and linear programming feasibility problems to make the interval observer operate with the ability to correctly estimate the system state with estimation errors within acceptable limits. The applicability of the proposed method is verified by simulation of the lateral vehicle control system.

ARTICLE HISTORY
Received 9 February 2023
Accepted 12 October 2023

KEYWORDS
Dynamical systems; neural networks; safety monitoring; interval observer

1. Introduction

Complex dynamical systems, such as autonomous vehicles and various cyber-physical systems (CPS), have been greatly benefiting from the fast advancement of artificial intelligence (AI) and machine learning (ML) technologies. Many new theories have been proposed on this basis, such as stable neural network controllers and observers (Levin & Narendra, 1992; Wu et al., 2014; L. Zhang et al., 2017), adaptive neural network controllers (Niu et al., 2020; Takahashi, 2017) and various neural network controllers (Hunt et al., 1992). Real-time monitoring of these dynamical systems embedded with neural network components is essential to ensure the system's safety. External inputs may have adversarial effects on the normal working state of the system; even with the most advanced neural networks, imperceptible perturbations in the input may lead to an erroneous result (Moosavi-Dezfooli et al., 2017). In addition, these systems are highly susceptible to erroneous outputs if they are subjected to adversarial attacks, which can have serious safety consequences. Therefore, to ensure the security of dynamical systems embedded in neural networks, it is essential to develop a technique that can monitor the operational state of dynamical systems in real time.

Most current approaches to safety or security verification take the form of offline computation. In general, verification using offline calculation requires a large amount of computational resources due to its high computational complexity. For example, for a type of neural networks with the activation function of rectified linear unit (ReLU), the safety verification problem can be represented as various complex computational problems. Based on polyhedral operations, a geometric computation method is proposed to obtain the exact output set of the neural network using ReLU activation function (Xiang et al., 2017b; Xiang, Tran, Rosenfeld et al., 2018). Based on those results, the methods in Tran, Manzanas Lopez et al. (2019) and Tran, Musau et al. (2019) extended it by proposing a novel approach with the aid of a specific convex set representation called star sets, which greatly improved scalability. A mixed-integer linear programming (MILP) method to validate neural networks was proposed in Lomuscio and Maganti (2017). The work (Dutta et al., 2019) focuses on neural networks with ReLU activation functions; they used a Taylor-model-based flowpipe construction scheme and replaced the neural network feedback with a polynomial mapping approach for a small fraction of the input to obtain an over-approximated reachable set. In addition, this method can be extended to other activation units after processing by segmental linearisation (Dutta et al., 2018). The work (Xiang, Tran & Johnson, 2018) introduces a simulation-based approach to output reachability estimation for neural networks with common activation functions. This paper (Xiang et al., 2021) takes the dynamic system embedded in the feedforward neural network named multilayer perceptrons (MLPs) as the research object, and develops a recursive algorithm with over-approximating the reachable set of the closed-loop system. The security verification of the system is achieved by checking the emptiness of the intersection between the insecure sets and the over-approximation of the reachable sets.

It is worth noting that the open-loop computational structure of these offline methods makes them quite challenging to implement in online settings. On the other hand, offline methods are difficult to detect system security issues in a timely

CONTACT Weiming Xiang ✉ wxiang@augusta.edu, xiangwming@gmail.com School of Computer and Cyber Sciences, Augusta University, Augusta, GA 30912, USA

manner, and the system state and parameters may differ from run-time when offline. Therefore, developing an online security monitoring method is very important. For this reason, inspired by observer design theory, we propose an alternative solution to design closed-loop systems for run-time monitoring based on instantaneous measurements of the system. We resort to develop interval observers for dynamical systems with neural networks. The interval observer can estimate the upper and lower bounds of the operating state trajectory of the dynamical system in real-time, which can achieve real-time safety monitoring of the dynamical system (Bolajraf et al., 2011; Cacace et al., 2015; Chebotarev et al., 2015; Efimov & Raïssi, 2016; Yw. Zhang et al., 2020). As shown in Xiang (2021), unlike the general interval observer design approach, the observer gains as well as auxiliary neural networks have to be designed through a series of optimisation problems to ensure that the interval observer can correctly estimate the upper and lower state bounds and a suitable estimation error. The design of the auxiliary neural networks in the interval observer is also necessary to simulate the behaviour of the neural network in the original system for better state estimation. The work (D. Zhang et al., 2020) applies interval observers to the safety monitoring of the state of charge (SOC) of lithium-ion batteries. The coupled equivalent circuit-thermal model is adopted in this paper, avoiding the complex structure and calculation caused by the traditional model with electrically and thermally coupled parallel connection of cells. The innovation of the work lies in considering cell heterogeneity as the uncertainty bounding functions and achieving the separation of the state number of interval observers from the number of parallel batteries.

During the design of interval observers, it is challenging to apply classical control theory, such as Lyapunov theory, for analysing dynamical systems embedded in neural network components due to the various types of nonlinear activation functions in neural networks. A popular approach is using quadratic constraints (QCs) to abstract the nonlinear activation functions in neural networks. The work (Anderson et al., 2007) analyses the stability of the feedback loop, including neural networks, by replacing the nonlinear and time-varying components of the neural networks with integral quadratic constraints (IQCs). Quadratic constraints are used to abstract the nonlinear activation functions and projection operators in neural network controllers in Hu et al. (2020), enabling the reachability analysis of closed-loop systems with neural network controllers. The approach in Fazlyab et al. (2022) uses quadratic constraints to abstract various properties of the activation function, such as bounded slope, monotonicity, and cross-layer repetition, thus formulating the safety verification problem for neural networks as the SDP feasibility problem. In addition, the characterisation of the input-output of neural networks through quadratic constraints allows other issues to be solved, such as the input-output sensitivity analysis of neural networks (Xiang, Tran & Johnson, 2018), safety verification and robustness analysis (Fazlyab et al., 2022), Lipschitz constant estimation of feedforward neural networks (Fazlyab et al., 2019), etc.

Synthesizing the previous discussions, the main contributions of this paper are as follows: (1) A global quadratic constraint formulation method for error dynamic systems is discussed; (2) A novel interval observer design method is proposed for the nonlinear dynamical systems with neural networks, and its core contribution is to abstract the nonlinear activation function of neural networks by the quadratic constraints method, so that some control theories applicable to linear systems can also be applied to the nonlinear dynamical systems with neural networks in this paper.

The rest of the paper is organised as follows. In Section 2, the system and problem formulation under discussion are presented. The main findings are given in Section 3, where the design methods for quadratic constraints on the activation function and auxiliary neural networks are presented, and the design of the interval observer gains $\underline{L}$ and $\overline{L}$ is represented in the form of a series of convex optimisation problems. The conclusion obtained is applied to a lateral control system for vehicles in Section 4. In Section 5, conclusions and future research directions are given.

Notations: In this paper, the notation $\mathbb{R}$ represents real numbers, and $\mathbb{R}_+$ is defined by $\mathbb{R}_+ = \{\tau \in \mathbb{R}, \tau \geq 0\}$. The notation $\mathbb{R}^n$ represents the vector space of all n-tuples of real numbers, and $\mathbb{R}^{n\times n}$ is the space of $n \times n$ matrices with real entries. The superscript 'T' denotes the matrix transpose. The block diagonal matrix is denoted by the symbol $diag\{\cdots\}$. The notation $I_n \in \mathbb{R}^{n\times n}$ denotes the n-dimensional identity matrix. Given a matrix $A \in \mathbb{R}^{m\times n}$, $\|A\|$ denote its Frobenius norm. For two vectors $x_1, x_2 \in \mathbb{R}^n$ or matrices $A_1, A_2 \in \mathbb{R}^{n\times n}$, the relations $x_1 < x_2$ and $A_1 < A_2$ are interpreted elementwisly. The relation $Q \succ 0$ $(Q \prec 0)$ means that $Q \in \mathbb{R}^{n\times n}$ is positive (negative) definite. In addition, $Q > 0$ $(Q \geq 0)$ means that all elements in this matrix $Q \in \mathbb{R}^{n\times n}$ are positive (nonnegative). $\mathbb{M}_n \in \mathbb{R}^{n\times n}$ is defined as the collection of all n-dimensional Metzler matrices.

2. System description and problem formulation

2.1 System description

In this paper, we consider a class of learning-enabled nonlinear dynamical systems embedded with neural networks in the following form

$$\begin{cases} \dot{x}(t) = f(x(t), u(t), \Phi(x(t))) \\ y(t) = g(x(t)) \end{cases}, \tag{1}$$

where $x \in \mathbb{R}^{n_x}$, $u(t) \in \mathbb{R}^{n_u}$ and $y \in \mathbb{R}^{n_y}$ are the state vector, input and output of the system, respectively. $f : \mathbb{R}^{n_x+n_u} \to \mathbb{R}^{n_x}$ and $g : \mathbb{R}^{n_x} \to \mathbb{R}^{n_y}$ are nonlinear functions. $\Phi : \mathbb{R}^{n_x} \to \mathbb{R}^{n_x}$ is the neural network component. Without causing ambiguity, we omit the time index t in some of the variables.

Specifically, this work considers a class of dynamical systems embedded with neural networks, which have the form of a Lipschitz nonlinear model as

$$\mathcal{L} : \begin{cases} \dot{x} = Ax + B_\Phi \Phi(x) + B_u u(t) + f(x) \\ y = Cx \end{cases}, \tag{2}$$

where $A \in \mathbb{R}^{n_x\times n_x}$, $B_\Phi \in \mathbb{R}_+^{n_x\times n_{L+1}}$, $B_u \in \mathbb{R}_+^{n_x\times n_u}$, $C \in \mathbb{R}^{n_y\times n_x}$ and $f(x)$ is a Lipschitz nonlinear function satisfying the following Lipschitz inequality

$$\|f(x_1) - f(x_2)\| \leq \beta\|x_1 - x_2\|, \beta > 0. \tag{3}$$

Remark 2.1: Many nonlinear systems in the form of $\dot{x} = f(x, u, \Phi(x))$ can be represented in the form of (2) if f is differentiable with respect to x and u. The neural network $\Phi(x)$ is the interval component that affects the behaviour of the system. For instance, the model (2) represents a state feedback closed-loop system if the neural network $\Phi(x)$ is trained as a feedback controller.

For the system (2), there are two sources of uncertainty: the initial values for state $x(0)$ and the instantaneous values of input $u(t)$. We assume that all these uncertainties belong to the known interval as shown in the following assumption.

Assumption 2.1: *Let $\underline{x}(0) \leq x(0) \leq \overline{x}(0)$ for some known $\underline{x}(0)$ and $\overline{x}(0) \in \mathbb{R}^{n_x}$, and let the known bounded functions $\underline{u}$ and $\overline{u}$ such that $\underline{u}(t) < u(t) < \overline{u}(t), \forall t \geq 0$.*

Suppose that the nonlinear function $f(x)$ has the following properties.

Assumption 2.2: *Suppose there exist functions $\underline{f}, \overline{f} : \mathbb{R}^{2n_x} \to \mathbb{R}^{n_x}$ such that*

$$\underline{f}(\underline{x}, \overline{x}) \leq f(x) \leq \overline{f}(\underline{x}, \overline{x}), \tag{4}$$

holds for any $\underline{x} \leq x \leq \overline{x}$.

Remark 2.2: Assumptions 2.1 and 2.2 emphasise that the initial state, the input signal and the nonlinear function of the original system, must numerically lie in the interval consisting of the initial state, the input signal and the nonlinear function of the interval observer, respectively. This is to ensure that the interval observer can correctly achieve the interval estimate for the state of the original system, which means $\underline{x} \leq x \leq \overline{x}$, in other words, to ensure that the error system is a positive system.

Assumption 2.3: *Suppose there exist scalars $\underline{a}_1, \overline{a}_1, \underline{a}_2, \overline{a}_2 \in \mathbb{R}_+$ and vectors $\underline{\rho}, \overline{\rho} \in \mathbb{R}_+^{n_x}$ such that*

$$f(x) - \underline{f}(\underline{x}, \overline{x}) \leq \underline{a}_1(x - \underline{x}) + \underline{a}_2(\overline{x} - x) + \underline{\rho},$$
$$\overline{f}(\underline{x}, \overline{x}) - f(x) \leq \overline{a}_1(x - \underline{x}) + \overline{a}_2(\overline{x} - x) + \overline{\rho},$$

holds for the nonlinear functions $\underline{f}(\underline{x}, \overline{x})$, $f(x)$, $\overline{f}(\underline{x}, \overline{x})$ defined in Assumption 2.2.

Remark 2.3: Under the Lipschitz condition (3), the estimation of parameters $\underline{a}_1$, $\overline{a}_1$, $\underline{a}_2$, $\overline{a}_2$, $\underline{\rho}$, $\overline{\rho}$ in Assumption 2.3 can be obtained through routine calculation, and the detailed estimation procedures can be found in Lemma 6 of Zheng et al. (2016)

An L-layer feedforward neural network $\Phi(x) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_{L+1}}$ is considered in this work, which is defined by the following recursive equation

$$\mathcal{N} : \begin{cases} \omega^{[0]} = x(t) & \\ \nu^{[l]} = W^{[l]}\omega^{[l-1]} + b^{[l]} & l = 1, \ldots, L \\ \omega^{[l]} = \phi^{[l]}(\nu^{[l]}) & l = 1, \ldots, L \\ \Phi(x) = W^{[L+1]}\omega^{[L]} + b^{[L+1]} & \end{cases}, \tag{5}$$

where $\omega^{[l]} \in \mathbb{R}^{n_l}$ denotes the output from the lth layer with n_l neurons of the neural network. $\nu^{[l]} \in \mathbb{R}^{n_l}$ denotes the input to the activation function of the lth layer of the neural network. $\Phi(x) \in \mathbb{R}^{n_{L+1}}$ is the output of the neural network feedback controller. $W^{[l]} \in \mathbb{R}^{n_l \times n_{l-1}}$ and $b^{[l]} \in \mathbb{R}^{n_l}$ represent the weight matrix and bias vector of the lth layer neural network, respectively. In the lth layer neural network, for vectors $\nu^{[l]} = [\nu_1^{[l]}, \nu_2^{[l]}, \ldots, \nu_{n_l}^{[l]}]^T$, we define $\phi^{[l]} = [\psi^{[l]}, \psi^{[l]}, \ldots, \psi^{[l]}]^T$ to be the series of activation functions and a single activation function is ψ, where $\phi^{[l]}(\nu^{[l]})$ is the action on each element in the vector, i.e.

$$\phi^{[l]}(\nu^{[l]}) = [\psi^{[l]}(\nu_1^{[l]}), \psi^{[l]}(\nu_2^{[l]}), \ldots, \psi^{[l]}(\nu_{n_l}^{[l]})]^T.$$

Here, the following assumptions about the activation function are given.

Assumption 2.4: *Suppose that for activation functions $\psi^{[l]}$, $l = 1, \ldots, L$, the following properties hold:*

- *Any two scalars x_1 and x_2 are given, there must be a scalar $\alpha > 0$ such that*

$$|\psi^{[l]}(x_1) - \psi^{[l]}(x_2)| \leq \alpha |x_1 - x_2|, \quad \forall l = 1, \ldots, L. \tag{6}$$

- *Any two scalars $x_1 \leq x_2$ are given, and we have*

$$\psi^{[l]}(x_1) \leq \psi^{[l]}(x_2), \quad \forall l = 1, \ldots, L. \tag{7}$$

Remark 2.4: Assumption 2.4 above applies to the most common activation functions, such as ReLU, sigmoid, tanh, and leaky ReLU. For condition (6), the α can be obtained by the maximum Lipschitz constant of all $\psi^{[l]}$. The condition (7) is satisfied because the common activation functions are monotonically increasing. Without loss of generality, we suppose that the activation functions are the same in each layer.

2.2 Problem formulation

Our proposed solution to the problem of safety monitoring of neural-network-embedded systems is to design a state estimator which is capable of estimating the upper and lower bounds of the state variable $x(t)$ to monitor the operation status of the system in real time. Information about the system $\mathcal{L}$ in the form of (2) being used for the estimator design includes: the system matrices A, B_Φ, B_u, C, the nonlinear function f, the neural network Φ, namely the weight matrix $\{W_l\}_{l=1}^{L+1}$, the bias vector $\{b_l\}_{l=1}^{L+1}$, the known bounded functions $\underline{u}, \overline{u}$ and the output $y(t)$. The run-time safety estimator design problem can be expressed as follows.

Problem 2.1: *For a dynamical system embedded with neural networks in the form of (2), how can we design a run-time safety state estimator such that its instantaneous state estimates, $\underline{x}$ and $\overline{x}$, satisfy $\underline{x} \leq x \leq \overline{x}, \forall t \geq 0$?*

To solve the above problem, we consider the development of a run-time safety state estimator in the form of the Luenberger interval observer

$$\begin{cases} \dot{\underline{x}} = (A - \underline{L}C)\underline{x} + \underline{L}y + B_\Phi \underline{\Phi}(\underline{x}, \overline{x}) + B_u \underline{u}(t) + \underline{f}(\underline{x}, \overline{x}) \\ \dot{\overline{x}} = (A - \overline{L}C)\overline{x} + \overline{L}y + B_\Phi \overline{\Phi}(\underline{x}, \overline{x}) + B_u \overline{u}(t) + \overline{f}(\underline{x}, \overline{x}) \end{cases} \tag{8}$$

where the initial state of the interval observer satisfies $\underline{x}(0) \leq x(0) \leq \overline{x}(0)$, $u(t)$ satisfies $\underline{u}(t) < u(t) < \overline{u}(t), \forall t \geq 0$, as shown in Assumption 2.1, and $\underline{f}(\underline{x}, \overline{x})$, $\overline{f}(\underline{x}, \overline{x})$ satisfy Assumptions 2.2 and 2.3. The auxiliary neural networks $\underline{\Phi}(\underline{x}, \overline{x})$ and $\overline{\Phi}(\underline{x}, \overline{x})$ and the observer gains $\underline{L}$ and $\overline{L}$ are to be determined.

Here, let the error state $\underline{e} = x - \underline{x}, \overline{e} = \overline{x} - x$, so that we can obtain the expression for the error dynamical system in the following form

$$\begin{cases} \dot{\underline{e}} = (A - \underline{L}C)\underline{e} + B_{\Phi}\Delta\underline{\Phi} + B_u(u - \underline{u}) + f(x) - \underline{f}(\underline{x}, \overline{x}) \\ \dot{\overline{e}} = (A - \overline{L}C)\overline{e} + B_{\Phi}\Delta\overline{\Phi} + B_u(\overline{u} - u) + \overline{f}(\underline{x}, \overline{x}) - f(x) \end{cases} \tag{9}$$

where $\Delta\underline{\Phi} = \Phi(x) - \underline{\Phi}(\underline{x}, \overline{x}), \Delta\overline{\Phi} = \overline{\Phi}(\underline{x}, \overline{x}) - \Phi(x)$, the initial state of the error system satisfy $\underline{e}(0) \geq 0$ and $\overline{e}(0) \geq 0$.

We find that the instantaneous estimates of the interval observer satisfy $\underline{x}(t) \leq x(t) \leq \overline{x}(t), \forall t \geq 0$ if we can make the state variable $\underline{e}(t) \geq 0, \overline{e}(t) \geq 0, \forall t \geq 0$. Thus, Problem 2.1 can be further formulated as follows.

Problem 2.2: *For a dynamical system embedded with neural networks in the form of* (2), *how can we design the observer gains* $\underline{L}$ *and* $\overline{L}$, *and the auxiliary neural networks* $\underline{\Phi}(\underline{x}, \overline{x})$ *and* $\overline{\Phi}(\underline{x}, \overline{x})$ *in the interval observer* (8) *such that error state instantaneous estimates* $\underline{e}(t) \geq 0$ *and* $\overline{e}(t) \geq 0, \forall t \geq 0$ *in error dynamical system* (9)*?*

To solve Problem 2.2, we review the conclusions related to positive systems.

Definition 2.1: If all elements outside the main diagonal of a matrix $A \in \mathbb{R}^{n \times n}$ are nonnegative, then $A \in \mathbb{M}_n$.

Lemma 2.1 (Wang et al., 2022): *The matrix* $PA \in \mathbb{M}_n$ *still holds if* P *is a diagonal positive definite matrix and* $A \in \mathbb{M}_n$.

Lemma 2.2 (Efimov & Raïssi, 2016): *Considering a system in the form of* $\dot{x}(t) = Ax(t) + d(t)$, *for* $A \in \mathbb{M}_n$, *the state* $x(t)$ *is elementwise nonnegative for all* $t \geq 0$ *if* $x(0) \geq 0$ *and* $d(t) \in \mathbb{R}^n_+$, *and the system is called cooperative.*

According to Lemma 2.2, we propose the following proposition as the solution to Problem 2.2, provided that $x(t)$ and $u(t)$ satisfy Assumption 2.1 and $\underline{f}(\underline{x}, \overline{x})$ and $\overline{f}(\underline{x}, \overline{x})$ satisfy Assumptions 2.2 and 2.3.

Proposition 2.1: *Problem* 2.2 *can be solved if the observer gains,* $\underline{L}$ *and* $\overline{L}$, *and the auxiliary neural networks,* $\underline{\Phi}(\underline{x}, \overline{x})$ *and* $\overline{\Phi}(\underline{x}, \overline{x})$, *satisfy the following conditions*

$$A - \underline{L}C \in \mathbb{M}_{n_x}, \tag{10}$$

$$A - \overline{L}C \in \mathbb{M}_{n_x}, \tag{11}$$

$$\Phi(x) - \underline{\Phi}(\underline{x}, \overline{x}) \in \mathbb{R}^{n_{L+1}}_+, \tag{12}$$

$$\overline{\Phi}(\underline{x}, \overline{x}) - \Phi(x) \in \mathbb{R}^{n_{L+1}}_+. \tag{13}$$

Proof: According to Assumptions 2.1, 2.2, it is clear that $f(x) - \underline{f}(\underline{x}, \overline{x}) \in \mathbb{R}^{n_x}_+$, $x(0) - \underline{x}(0) \in \mathbb{R}^{n_x}_+$ and $B_u(u - \underline{u}) \in \mathbb{R}^{n_x}_+$. Since $B_{\Phi}(\Phi(x) - \underline{\Phi}(\underline{x}, \overline{x})) \in \mathbb{R}^{n_x}_+$ holds and $A - \underline{L}C \in \mathbb{M}_{n_x}$, according to Lemma 2.2, we can conclude $\underline{e}(t) \geq 0, \forall t \geq 0$. The same can be said for $\overline{e}(t) \geq 0, \forall t \geq 0$. Thus the proof is complete. ■

It is worth noting that the conditions in Proposition 2.1 hold only to prove that $\underline{e}(t) \geq 0, \overline{e}(t) \geq 0, \forall t \geq 0$. Under the conditions that Proposition 2.1 holds, it is possible that $\lim_{t \to \infty} \underline{e}(t) = \infty$ and $\lim_{t \to \infty} \overline{e}(t) = \infty$ happen. Although the interval observer (8) can provide estimated boundaries of the states of the system (2), the estimation error can be extremely large making the estimates meaningless. Therefore, the concept of practical stability, which is related to the boundedness of the system states as time grows, is introduced.

Lemma 2.3 (Ge & Wang, 2004): *Considering the system* (2), *if there exists a continuous Lyapunov function* $V(x)$ *satisfying* $a_1(\| x \|) \leq V(x) \leq a_2(\| x \|)$, *making* $\dot{V}(x) \leq -c_1 V(x) + c_2$, *where* a_1 *and* a_2 *are class* $\mathcal{K}$ *functions of the state* x, *and* c_1 *and* c_2 *are positive constants, then the solution* $x(t)$ *is uniformly bounded and the system is globally practically uniformly exponentially stable.*

3. Observer-based safety monitoring design

The aim of this section is to design the interval observer gains $\underline{L}$ and $\overline{L}$, and the auxiliary neural networks $\underline{\Phi}(\underline{x}, \overline{x})$ and $\overline{\Phi}(\underline{x}, \overline{x})$ that satisfy Proposition 2.1. In order to minimise the estimation errors, the convergence of the error system also needs to be considered. First, we introduce the design method of auxiliary neural networks $\underline{\Phi}(\underline{x}, \overline{x})$ and $\overline{\Phi}(\underline{x}, \overline{x})$ based on the neural network $\Phi(x)$ defined in (5).

For a given neural network Φ, the lth layer weight matrix is in the following form of

$$W^{[l]} = [w^{[l]}_{i,j}] = \begin{bmatrix} w^{[l]}_{1,1} & w^{[l]}_{1,2} & \cdots & w^{[l]}_{1,n_{l-1}} \\ w^{[l]}_{2,1} & w^{[l]}_{2,2} & \cdots & w^{[l]}_{2,n_{l-1}} \\ \vdots & \vdots & \ddots & \vdots \\ w^{[l]}_{n_l,1} & w^{[l]}_{n_l,2} & \cdots & w^{[l]}_{n_l,n_{l-1}} \end{bmatrix}, \tag{14}$$

where $w^{[l]}_{i,j}$ expresses the element in ith row and jth column. Two auxiliary weight matrices are defined as follows

$$\begin{aligned} \underline{W}^{[l]} &= [\underline{w}^{[l]}_{i,j}], \underline{w}^{[l]}_{i,j} = \begin{cases} w^{[l]}_{i,j}, & w^{[l]}_{i,j} < 0 \\ 0, & w^{[l]}_{i,j} \geq 0 \end{cases}, \\ \overline{W}^{[l]} &= [\overline{w}^{[l]}_{i,j}], \overline{w}^{[l]}_{i,j} = \begin{cases} w^{[l]}_{i,j}, & w^{[l]}_{i,j} \geq 0 \\ 0, & w^{[l]}_{i,j} < 0 \end{cases}. \end{aligned} \tag{15}$$

Obviously, we can get $W^{[l]} = \underline{W}^{[l]} + \overline{W}^{[l]}$. Then two auxiliary neural networks $\underline{\Phi}(\underline{x}, \overline{x}) : \mathbb{R}^{2n_0} \to \mathbb{R}^{n_{L+1}}$ and $\overline{\Phi}(\underline{x}, \overline{x}) : \mathbb{R}^{2n_0} \to \mathbb{R}^{n_{L+1}}$ are constructed with inputs $\underline{x}, \overline{x} \in \mathbb{R}^{n_0}$ in the expression of

$$\underline{\mathcal{N}} : \begin{cases} \underline{\omega}^{[0]} = \underline{x}(t) \\ \underline{v}^{[l]} = \underline{W}^{[l]}\overline{\omega}^{[l-1]} + \overline{W}^{[l]}\underline{\omega}^{[l-1]} + b^{[l]} \\ \underline{\omega}^{[l]} = \phi^{[l]}(\underline{v}^{[l]}) \\ \underline{\Phi}(\underline{x}, \overline{x}) = \underline{W}^{[L+1]}\overline{\omega}^{[L]} + \overline{W}^{[L+1]}\underline{\omega}^{[L]} + b^{[L+1]} \end{cases}, \tag{16}$$

$$\overline{\mathcal{N}}:\begin{cases}\overline{\omega}^{[0]} = \overline{x}(t)\\ \overline{v}^{[l]} = \underline{W}^{[l]}\underline{\omega}^{[l-1]} + \overline{W}^{[l]}\overline{\omega}^{[l-1]} + b^{[l]}\\ \overline{\omega}^{[l]} = \phi^{[l]}(\overline{v}^{[l]})\\ \overline{\Phi}(\underline{x},\overline{x}) = \underline{W}^{[L+1]}\underline{\omega}^{[L]} + \overline{W}^{[L+1]}\overline{\omega}^{[L]} + b^{[L+1]}\end{cases}, \tag{17}$$

where $l = 1,\ldots,L$.

In the case $\underline{x} \le x \le \overline{x}$, the following lemma proves that the auxiliary neural networks $\underline{\Phi}(\underline{x},\overline{x})$ and $\overline{\Phi}(\underline{x},\overline{x})$ identified by (16) and (17) can satisfy (12) and (13) in Proposition 2.1, i.e. $\Phi(x) - \underline{\Phi}(\underline{x},\overline{x}) \in \mathbb{R}_+^{n_{L+1}}$, $\overline{\Phi}(\underline{x},\overline{x}) - \Phi(x) \in \mathbb{R}_+^{n_{L+1}}$.

Lemma 3.1 (Xiang, 2021): *Considering the neural network* $\Phi : \mathbb{R}^{n_0} \to \mathbb{R}^{n_{L+1}}$ *and auxiliary neural networks* $\underline{\Phi}(\underline{x},\overline{x}) : \mathbb{R}^{2n_0} \to \mathbb{R}^{n_{L+1}}$, $\overline{\Phi}(\underline{x},\overline{x}) : \mathbb{R}^{2n_0} \to \mathbb{R}^{n_{L+1}}$ *described by* (16) *and* (17), *the following condition*

$$\begin{bmatrix}\Phi(x) - \underline{\Phi}(\underline{x},\overline{x})\\ \overline{\Phi}(\underline{x},\overline{x}) - \Phi(x)\end{bmatrix} \in \mathbb{R}_+^{2n_{L+1}}, \tag{18}$$

holds for any $\underline{x} \le x \le \overline{x}$.

The above constructed neural networks and Lemma 3.1 provide a method for designing the auxiliary neural networks $\underline{\Phi}(\underline{x},\overline{x})$ and $\overline{\Phi}(\underline{x},\overline{x})$ that meet the conditions in Proposition 2.1. Next, we need to design the observer gains $\underline{L}$ and $\overline{L}$ such that (10) and (11) in Proposition 2.1 hold and the estimation error is within an acceptable range. The nonlinear activation function makes it difficult to incorporate the above results into the convex optimisation framework which is usually used for observer gain design. Inspired by the approach proposed in the literature (Yin et al., 2022), we can abstract the activation function by quadratic constraints.

3.1 Quadratic constraints on the activation functions

Considering the error dynamical system (9) and in connection with the definition of the auxiliary neural networks (16) and (17), the following results can be obtained

$$\begin{aligned}\Phi - \underline{\Phi} &= W^{[L+1]}\omega^{[L]} + b^{[L+1]} - (\underline{W}^{[L+1]}\overline{\omega}^{[L]}\\ &\quad + \overline{W}^{[L+1]}\underline{\omega}^{[L]} + b^{[L+1]})\\ &= (\underline{W}^{[L+1]} + \overline{W}^{[L+1]})\omega^{[L]} - (\underline{W}^{[L+1]}\overline{\omega}^{[L]}\\ &\quad + \overline{W}^{[L+1]}\underline{\omega}^{[L]})\\ &= \overline{W}^{[L+1]}\underline{\xi}^{[L]} - \underline{W}^{[L+1]}\overline{\xi}^{[L]},\\ \overline{\Phi} - \Phi &= \underline{W}^{[L+1]}\underline{\omega}^{[L]} + \overline{W}^{[L+1]}\overline{\omega}^{[L]} + b^{[L+1]}\\ &\quad - (W^{[L+1]}\omega^{[L]} + b^{[L+1]})\\ &= \underline{W}^{[L+1]}\underline{\omega}^{[L]} + \overline{W}^{[L+1]}\overline{\omega}^{[L]} - (\underline{W}^{[L+1]}\\ &\quad + \overline{W}^{[L+1]})\omega^{[L]}\\ &= -\underline{W}^{[L+1]}\underline{\xi}^{[L]} + \overline{W}^{[L+1]}\overline{\xi}^{[L]},\end{aligned}$$

$$v_\Phi - \underline{v}_\Phi = \begin{bmatrix} v^{[1]} - \underline{v}^{[1]}\\ v^{[2]} - \underline{v}^{[2]}\\ \vdots\\ v^{[L]} - \underline{v}^{[L]}\end{bmatrix} = \begin{bmatrix}\overline{W}^{[1]}(x - \underline{x}) - \underline{W}^{[1]}(\overline{x} - x)\\ \overline{W}^{[2]}\underline{\xi}^{[1]} - \underline{W}^{[2]}\overline{\xi}^{[1]}\\ \vdots\\ \overline{W}^{[L]}\underline{\xi}^{[L-1]} - \underline{W}^{[L]}\overline{\xi}^{[L-1]}\end{bmatrix},$$

$$\overline{v}_\Phi - v_\Phi = \begin{bmatrix}\overline{v}^{[1]} - v^{[1]}\\ \overline{v}^{[2]} - v^{[2]}\\ \vdots\\ \overline{v}^{[L]} - v^{[L]}\end{bmatrix} = \begin{bmatrix}-\underline{W}^{[1]}(x - \underline{x}) + \overline{W}^{[1]}(\overline{x} - x)\\ -\underline{W}^{[2]}\underline{\xi}^{[1]} + \overline{W}^{[2]}\overline{\xi}^{[1]}\\ \vdots\\ -\underline{W}^{[L]}\underline{\xi}^{[L-1]} + \overline{W}^{[L]}\overline{\xi}^{[L-1]}\end{bmatrix},$$

where $\underline{\xi}^{[l]} = \omega^{[l]} - \underline{\omega}^{[l]}$ and $\overline{\xi}^{[l]} = \overline{\omega}^{[l]} - \omega^{[l]}$.

Furthermore, the following relationship is readily available

$$\begin{bmatrix}\Phi - \underline{\Phi}\\ \overline{\Phi} - \Phi\\ v_\Phi - \underline{v}_\Phi\\ \overline{v}_\Phi - v_\Phi\end{bmatrix} = N\begin{bmatrix}x - \underline{x}\\ \overline{x} - x\\ \omega_\Phi - \underline{\omega}_\Phi\\ \overline{\omega}_\Phi - \omega_\Phi\end{bmatrix}, \tag{19}$$

where N is defined in Table 1, and

$$\underline{\omega}_\Phi(t) = \begin{bmatrix}\underline{\omega}^{[1]}(t)\\ \vdots\\ \underline{\omega}^{[L]}(t)\end{bmatrix},\quad \overline{\omega}_\Phi(t) = \begin{bmatrix}\overline{\omega}^{[1]}(t)\\ \vdots\\ \overline{\omega}^{[L]}(t)\end{bmatrix},$$

$$\phi(\underline{v}_\Phi) = \begin{bmatrix}\phi^{[1]}(\underline{v}^{[1]})\\ \vdots\\ \phi^{[L]}(\underline{v}^{[L]})\end{bmatrix} \in \mathbb{R}^{n_\Phi},$$

$$\phi(\overline{v}_\Phi) = \begin{bmatrix}\phi^{[1]}(\overline{v}^{[1]})\\ \vdots\\ \phi^{[L]}(\overline{v}^{[L]})\end{bmatrix} \in \mathbb{R}^{n_\Phi},$$

in which $n_\Phi = n_1 + n_2 + \cdots + n_L$.

Abstracting the activation function based on quadratic constraints (QCs) is an essential approach in the following interval observer design. Let us first define an offset local sector.

Definition 3.1 (Yin et al., 2022): Suppose that given $\alpha, \beta, \hat{\underline{v}}, \hat{\overline{v}}, v^* \in \mathbb{R}$, where $\alpha \le \beta, \hat{\underline{v}} \le v^* \le \hat{\overline{v}}$. The activation function $\psi : \mathbb{R} \to \mathbb{R}$ satisfies the offset local sector $[\alpha, \beta]$ around the given point $(v^*, \psi(v^*))$ if

$$(\Delta\psi(v) - \alpha\Delta v)(\beta\Delta v - \Delta\psi(v)) \ge 0, \quad \forall v \in [\hat{\underline{v}}, \hat{\overline{v}}], \tag{20}$$

where $\Delta v = v - v^*$ and $\Delta\psi(v) = \psi(v) - \psi(v^*)$.

If the function ψ satisfies a local offset sector $[\alpha, \beta]$ centred at any point $(v^*, \psi(v^*))$, it means that the function ψ satisfies a global offset sector $[\alpha, \beta]$. As shown in Figure 1(a), function $\psi(v) = tanh(v)$ satisfies the global sector bound around the point $(1, \psi(1))$ with $[\alpha, \beta] = [0, 1]$. For global sector constraints, the value of α, β are independent of the chosen reference point $(v^*, \psi(v^*))$ and are only related to the chosen activation function. When the input to the function is restricted to $v \in [\hat{\underline{v}}, \hat{\overline{v}}]$, the stricter offset local sector constraint will be

Table 1. Definition of N in (19).

$$N=\left[\begin{array}{c|c|c|c}
\overline{N}_{\Phi x} & \underline{N}_{\Phi x} & \overline{N}_{\Phi\omega} & \underline{N}_{\Phi\omega}\\ \hline
\underline{N}_{\Phi x} & \overline{N}_{\Phi x} & \underline{N}_{\Phi\omega} & \overline{N}_{\Phi\omega}\\ \hline
\overline{N}_{vx} & \underline{N}_{vx} & \overline{N}_{v\omega} & \underline{N}_{v\omega}\\ \hline
\underline{N}_{vx} & \overline{N}_{vx} & \underline{N}_{v\omega} & \overline{N}_{v\omega}
\end{array}\right]$$

$$=\left[\begin{array}{c|c|cccc|cccc}
0 & 0 & 0 & 0 & \cdots & \overline{W}^{[L+1]} & 0 & 0 & \cdots & -\underline{W}^{[L+1]}\\ \hline
0 & 0 & 0 & 0 & \cdots & -\underline{W}^{[L+1]} & 0 & 0 & \cdots & \overline{W}^{[L+1]}\\ \hline
\overline{W}^{[1]} & -\underline{W}^{[1]} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0\\
0 & 0 & \overline{W}^{[2]} & \cdots & 0 & 0 & -\underline{W}^{[2]} & \cdots & 0 & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\
0 & 0 & 0 & \cdots & \overline{W}^{[L]} & 0 & 0 & \cdots & -\underline{W}^{[L]} & 0\\ \hline
-\underline{W}^{[1]} & \overline{W}^{[1]} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0\\
0 & 0 & -\underline{W}^{[2]} & \cdots & 0 & 0 & \overline{W}^{[2]} & \cdots & 0 & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\
0 & 0 & 0 & \cdots & -\underline{W}^{[L]} & 0 & 0 & \cdots & \overline{W}^{[L]} & 0
\end{array}\right]$$

satisfied. As shown in Figure 1(b), function $\psi(v)=tanh(v)$ satisfies the offset local sector bound around the point $(0,\psi(0))$ with $[\alpha,\beta]=[0.48,1]$, where $v\in[-2,2]$.

We can then convert the expression (20) of the local offset sector into the following form

$$\alpha\le\frac{\psi(v)-\psi(v^*)}{v-v^*}\le\beta,\quad \forall v\in[\underline{\hat{v}},\hat{\overline{v}}]. \tag{21}$$

According to (21), we can further interpret Definition 3.1 as follows. For a function ψ satisfying a local offset sector $[\alpha,\beta]$ around the point $(v^*,\psi(v^*))$, considering $\forall v\in[\underline{\hat{v}},\hat{\overline{v}}]$, the slope of the line connecting any point on the function ψ to the centre point $(v^*,\psi(v^*))$ is between $[\alpha,\beta]$. The local sector constraint for a single activation function $\psi:\mathbb{R}\to\mathbb{R}$ is given above. Next, we consider the local sector constraint problem for a function formed by concatenating multiple activation functions. Considering the activation function of a series connection $\phi^{n_\Phi}:\mathbb{R}^{n_\Phi}\to\mathbb{R}^{n_\Phi}$, given $\alpha_\Phi,\beta_\Phi,\underline{\hat{v}}_\Phi,\hat{\overline{v}}_\Phi,v^*_\Phi\in\mathbb{R}^{n_\Phi}$, satisfying $\alpha_\Phi\le\beta_\Phi$, $\underline{\hat{v}}_\Phi\le v^*_\Phi\le\hat{\overline{v}}_\Phi$, for the ith input $v_{\Phi,i}\in[\underline{\hat{v}}_{\Phi,i},\hat{\overline{v}}_{\Phi,i}]$, $i=1,\dots,n_\Phi$ of the function ϕ^{n_Φ}, we can obtain the offset sector $[\alpha_{\Phi,i},\beta_{\Phi,i}]$ either analytically or numerically. α_Φ,β_Φ can be obtained by stacking these local sectors, and the quadratic constraints considering the concatenation of activation functions ϕ^{n_Φ} is given below.

Lemma 3.2 (Yin et al., 2022): *Given $\alpha_\Phi,\beta_\Phi,\underline{\hat{v}}_\Phi,\hat{\overline{v}}_\Phi,v^*_\Phi\in\mathbb{R}^{n_\Phi}$, satisfying $\alpha_\Phi\le\beta_\Phi$, $\underline{\hat{v}}_\Phi\le v^*_\Phi\le\hat{\overline{v}}_\Phi$ and $\omega^*_\Phi=\phi(v^*_\Phi)$. Suppose that the function $\phi^{n_\Phi}:\mathbb{R}^{n_\Phi}\to\mathbb{R}^{n_\Phi}$ satisfies the offset local sector $[\alpha_\Phi,\beta_\Phi]$ around the point $(v^*_\Phi,\psi(v^*_\Phi))$. Given $\lambda\ge0$ where $\lambda\in\mathbb{R}^{n_\Phi}$, we have*

$$\begin{bmatrix} v_\Phi-v^*_\Phi\\ \omega_\Phi-\omega^*_\Phi\end{bmatrix}^T\tilde{\Psi}^T_\Phi\tilde{M}_\Phi(\lambda)\tilde{\Psi}_\Phi\begin{bmatrix} v_\Phi-v^*_\Phi\\ \omega_\Phi-\omega^*_\Phi\end{bmatrix}\ge0, \tag{22}$$

where $\omega_\Phi=\phi^{n_\Phi}(v)$ and

$$\tilde{\Psi}_\Phi=\begin{bmatrix} diag(\beta_\Phi) & -I_{n_\Phi}\\ -diag(\alpha_\Phi) & I_{n_\Phi}\end{bmatrix},$$

$$\tilde{M}_\Phi(\lambda)=\begin{bmatrix} 0_{n_\Phi} & diag(\lambda)\\ diag(\lambda) & 0_{n_\Phi}\end{bmatrix}.$$

Lemma 3.2 considers the problem of quadratic constraints on the local offset sector at the level of the activation function of the entire neural network. Since our interval observers need to work properly for any input, it is necessary to consider quadratic constraints on the activation function for the global sector. Let $\underline{\hat{v}}\to-\infty,\hat{\overline{v}}\to\infty$, considering the activation function $\psi(v)=tanh(v)$, then (22) holds if $\alpha=0,\beta=1$. According to Definition 3.1, $v^*\in\mathbb{R}$ in this case, it is feasible that $v^*=\underline{v}^{[l]}_i$ or $v^*=\overline{v}^{[l]}_i, l=1,\dots,L, i=1,\dots,n_l$. Thus, we can get

$$\alpha\le\frac{\psi^{[l]}(v^{[l]}_i)-\psi^{[l]}(\underline{v}^{[l]}_i)}{v^{[l]}_i-\underline{v}^{[l]}_i}\le\beta, \tag{23}$$

$$\alpha\le\frac{\psi^{[l]}(\overline{v}^{[l]}_i)-\psi^{[l]}(v^{[l]}_i)}{\overline{v}^{[l]}_i-v^{[l]}_i}\le\beta. \tag{24}$$

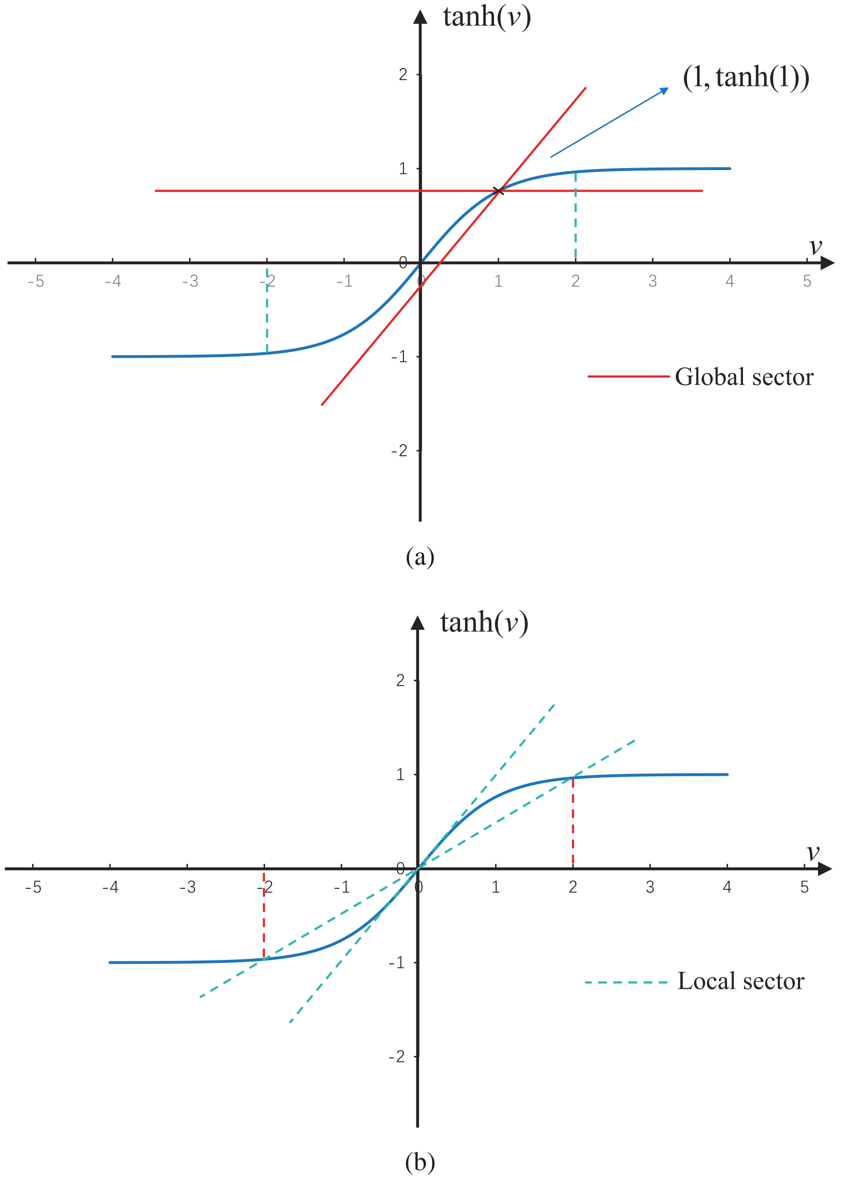
Similarly, considering the case for the global sector, we can obtain the global sector quadratic constraints on the activation function applied to the neural network (5) and the auxiliary neural networks (16) and (17) in the error dynamical system (9) as follows.

Theorem 3.1: *Given $\alpha_\Phi,\beta_\Phi\in\mathbb{R}^{n_\Phi}$ and existing $\underline{v}_\Phi,\overline{v}_\Phi,v_\Phi\in\mathbb{R}^{n_\Phi}$, satisfying $\alpha_\Phi\le\beta_\Phi$, $\underline{v}_\Phi\le v_\Phi\le\overline{v}_\Phi$ and $\omega_\Phi=\phi^{n_\Phi}(v_\Phi)$. Consider the definition of the neural network (5) and auxiliary neural networks (16) and (17), for exactly the same activation function of the concatenation $\phi^{n_\Phi}=[\psi,\dots,\psi]:\mathbb{R}^{n_\Phi}\to\mathbb{R}^{n_\Phi}$. Given $\lambda\ge0$ where $\lambda\in\mathbb{R}^{n_\Phi}$, we have*

$$\Pi=\begin{bmatrix} v_\Phi-\underline{v}_\Phi\\ \overline{v}_\Phi-v_\Phi\\ \omega_\Phi-\underline{\omega}_\Phi\\ \overline{\omega}_\Phi-\omega_\Phi\end{bmatrix}^T\Psi^T_\Phi M_\Phi(\lambda)\Psi_\Phi\begin{bmatrix} v_\Phi-\underline{v}_\Phi\\ \overline{v}_\Phi-v_\Phi\\ \omega_\Phi-\underline{\omega}_\Phi\\ \overline{\omega}_\Phi-\omega_\Phi\end{bmatrix}\ge0, \tag{25}$$

where

$$\Psi_\Phi=\begin{bmatrix} diag(\beta_\Phi) & 0_{n_\Phi} & -I_{n_\Phi} & 0_{n_\Phi}\\ 0_{n_\Phi} & diag(\beta_\Phi) & 0_{n_\Phi} & -I_{n_\Phi}\\ -diag(\alpha_\Phi) & 0_{n_\Phi} & I_{n_\Phi} & 0_{n_\Phi}\\ 0_{n_\Phi} & -diag(\alpha_\Phi) & 0_{n_\Phi} & I_{n_\Phi}\end{bmatrix},$$



(a)

(b)

Figure 1. Two types of quadratic constraints. (a) Global sector constraint on function $\psi(v) = tanh(v)$ and (b) Offset local sector constraint on function $\psi(v) = tanh(v)$.

$$M_\Phi(\lambda) = \begin{bmatrix} 0_{n_\Phi} & 0_{n_\Phi} & diag(\lambda) & 0_{n_\Phi} \\ 0_{n_\Phi} & 0_{n_\Phi} & 0_{n_\Phi} & diag(\lambda) \\ diag(\lambda) & 0_{n_\Phi} & 0_{n_\Phi} & 0_{n_\Phi} \\ 0_{n_\Phi} & diag(\lambda) & 0_{n_\Phi} & 0_{n_\Phi} \end{bmatrix},$$

and $\alpha_\Phi = [\alpha, \dots, \alpha]^T, \beta_\Phi = [\beta, \dots, \beta]^T \in \mathbb{R}^{n_\Phi}$, *which can be obtained by analysing the global sector constraints on a single activation function.*

Proof: According to (25), for any $\underline{v}_\Phi \le v_\Phi \le \bar{v}_\Phi$, one can obtain

$$\begin{bmatrix} v_\Phi - \underline{v}_\Phi \\ \bar{v}_\Phi - v_\Phi \\ \omega_\Phi - \underline{\omega}_\Phi \\ \bar{\omega}_\Phi - \omega_\Phi \end{bmatrix}^T \Psi_\Phi^T M_\Phi(\lambda) \Psi_\Phi \begin{bmatrix} v_\Phi - \underline{v}_\Phi \\ \bar{v}_\Phi - v_\Phi \\ \omega_\Phi - \underline{\omega}_\Phi \\ \bar{\omega}_\Phi - \omega_\Phi \end{bmatrix}$$

$$= \begin{bmatrix} \Delta\underline{v}_\Phi \\ \Delta\bar{v}_\Phi \\ \Delta\underline{\omega}_\Phi \\ \Delta\bar{\omega}_\Phi \end{bmatrix}^T \Psi_\Phi^T M_\Phi(\lambda) \Psi_\Phi \begin{bmatrix} \Delta\underline{v}_\Phi \\ \Delta\bar{v}_\Phi \\ \Delta\underline{\omega}_\Phi \\ \Delta\bar{\omega}_\Phi \end{bmatrix}$$

$$= \sum_{l=1}^{L} \sum_{i=1}^{n_l} \lambda_i^{[l]} (\Delta\underline{\omega}_i^{[l]} - \alpha\Delta\underline{v}_i^{[l]})(\beta\Delta\underline{v}_i^{[l]} - \Delta\underline{\omega}_i^{[l]})$$

$$+ \sum_{l=1}^{L} \sum_{i=1}^{n_l} \lambda_i^{[l]} (\Delta\bar{\omega}_i^{[l]} - \alpha\Delta\bar{v}_i^{[l]})(\beta\Delta\bar{v}_i^{[l]} - \Delta\bar{\omega}_i^{[l]}),$$

where $\Delta\underline{v}_i^{[l]} = v_i^{[l]} - \underline{v}_i^{[l]}$, $\Delta\bar{v}_i^{[l]} = \bar{v}_i^{[l]} - v_i^{[l]}$, $\Delta\underline{\omega}_i^{[l]} = \omega_i^{[l]} - \underline{\omega}_i^{[l]}$, $\Delta\bar{\omega}_i^{[l]} = \bar{\omega}_i^{[l]} - \omega_i^{[l]}$. Using (23) and (24), it is easy to see that each term in the equation is non-negative in the case of $\lambda_i^{[l]} \ge 0$. Thus the proof is complete. ■

Remark 3.1: For the description of the quadratic constraints on activation functions, our approach uses an extension based on the description of the local constraint in Definition 3.1 to obtain the global constraint needed for the subsequent proof. For example, for the activation function $\psi(v) = tanh(v)$ mentioned in Figure 1(a), we have $\alpha = 0$, $\beta = 1$. Therefore, we can obtain $\alpha_\Phi = [0, \dots, 0]^T, \beta_\Phi = [1, \dots, 1]^T$. Definition 3.1 can be generalised from local constraint to global constraint, and a more detailed discussion on the quadratic constraints can be found in the paper (Yin et al., 2022).

3.2 Design of interval observer

This section uses the Lyapunov stability theory and the global sector constraint of the activation function given in Theorem 3.1 to obtain the tractable linear matrix inequality (LMI) and conditions that ensure the error dynamical system (9) a positive system and practically stable.

Theorem 3.2: *Considering the error dynamical system* (9) *with the definition of neural network* (5) *and the definition of auxiliary neural networks* (16) *and* (17), *and nonlinear function under Assumptions 2.2, 2.3, if there exist diagonal matrix $Q \succ 0$, diagonal matrix S and block diagonal matrix M, real number $k_1 > 0$, such that*

$$\begin{bmatrix} \Gamma_1 & \Gamma_2 & Q\tilde{B}_\Phi & Q\tilde{B}_u & Q \\ \Gamma_2^T & \Gamma_3 & 0 & 0 & 0 \\ \tilde{B}_\Phi^T Q & 0 & -I & 0 & 0 \\ \tilde{B}_u^T Q & 0 & 0 & -I & 0 \\ Q & 0 & 0 & 0 & -I \end{bmatrix} \prec 0, \tag{26}$$

$$Q\tilde{A} - M\tilde{C} + S \geq 0, \tag{27}$$

where $\Gamma_1 = Q\tilde{A} - M\tilde{C} + \tilde{A}^T Q - \tilde{C}^T M^T + k_1 I + \tilde{N}_{\Phi x}^T \tilde{N}_{\Phi x} + \tilde{N}_{vx}^T F_{\alpha\beta} \tilde{N}_{vx}$, $\Gamma_2 = \tilde{N}_{\Phi x}^T \tilde{N}_{\Phi\omega} + \tilde{N}_{vx}^T F_{\alpha\beta} \tilde{N}_{v\omega} + \tilde{N}_{vx}^T F_{\alpha+\beta}$, *and* $\Gamma_3 = \tilde{N}_{\Phi\omega}^T \tilde{N}_{\Phi\omega} + \tilde{N}_{v\omega}^T F_{\alpha\beta} \tilde{N}_{v\omega} + F_{\alpha+\beta} \tilde{N}_{v\omega} + \tilde{N}_{v\omega}^T F_{\alpha+\beta} + F_\lambda$ *in which*

$$\tilde{A} = \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix}, \quad \tilde{B}_\Phi = \begin{bmatrix} B_\Phi & 0 \\ 0 & B_\Phi \end{bmatrix}, \quad \tilde{B}_u = \begin{bmatrix} B_u & 0 \\ 0 & B_u \end{bmatrix},$$

$$\tilde{C} = \begin{bmatrix} C & 0 \\ 0 & C \end{bmatrix}, \quad \tilde{L} = \begin{bmatrix} \underline{L} & 0 \\ 0 & \overline{L} \end{bmatrix}, \quad M = \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix},$$

$$\begin{bmatrix} \tilde{N}_{\Phi x} & \tilde{N}_{\Phi\omega} \\ \tilde{N}_{vx} & \tilde{N}_{v\omega} \end{bmatrix} = \left[\begin{array}{cc:cc} \overline{N}_{\Phi x} & \underline{\underline{N}}_{\Phi x} & \overline{N}_{\Phi\omega} & \underline{\underline{N}}_{\Phi\omega} \\ \underline{N}_{\Phi x} & \overline{\overline{N}}_{\Phi x} & \underline{N}_{\Phi\omega} & \overline{\overline{N}}_{\Phi\omega} \\ \hdashline \overline{N}_{vx} & \underline{\underline{N}}_{vx} & \overline{N}_{v\omega} & \underline{\underline{N}}_{v\omega} \\ \underline{N}_{vx} & \overline{\overline{N}}_{vx} & \underline{N}_{v\omega} & \overline{\overline{N}}_{v\omega} \end{array}\right],$$

$$\Psi_\Phi^T M_\Phi(\lambda)\Psi_\Phi = \begin{bmatrix} F_{\alpha\beta} & F_{\alpha+\beta} \\ F_{\alpha+\beta} & F_\lambda \end{bmatrix}$$

$$= \left[\begin{array}{cc:cc} \lambda_1 & 0_{n_\Phi} & \lambda_2 & 0_{n_\Phi} \\ 0_{n_\Phi} & \lambda_1 & 0_{n_\Phi} & \lambda_2 \\ \hdashline \lambda_2 & 0_{n_\Phi} & \lambda_3 & 0_{n_\Phi} \\ 0_{n_\Phi} & \lambda_2 & 0_{n_\Phi} & \lambda_3 \end{array}\right],$$

$$\lambda_1 = -2\alpha\beta diag(\lambda), \lambda_2 = (\alpha+\beta) diag(\lambda), \lambda_3 = -2diag(\lambda),$$

and $\alpha, \beta \in \mathbb{R}$ are the exact values determined by the chosen activation function, and $k_1 = 3max\{(\underline{a}_1^2 + \overline{a}_1^2), (\underline{a}_2^2 + \overline{a}_2^2)\}$, which can be calculated by Assumption 2.3, then the error dynamical system (9) *is a practically stable and positive system. The system* (8) *is an interval observer of the nonlinear system* (2) *and the observer gains matrices $\underline{L}$ and $\overline{L}$ can be obtained by $\tilde{L} = Q^{-1}M$.*

Proof: Since (26) holds, let $E = \tilde{A} - \tilde{L}\tilde{C}$ and according to the relation $\tilde{L} = Q^{-1}M$, (26) can be rewritten as

$$\begin{bmatrix} \Gamma_4 & \Gamma_2 & Q\tilde{B}_\Phi & Q\tilde{B}_u & Q \\ \Gamma_2^T & \Gamma_3 & 0 & 0 & 0 \\ \tilde{B}_\Phi^T Q & 0 & -I & 0 & 0 \\ \tilde{B}_u^T Q & 0 & 0 & -I & 0 \\ Q & 0 & 0 & 0 & -I \end{bmatrix} \prec 0, \tag{28}$$

where $\Gamma_4 = QE + E^T Q + k_1 I + \tilde{N}_{\Phi x}^T \tilde{N}_{\Phi x} + \tilde{N}_{vx}^T F_{\alpha\beta} \tilde{N}_{vx}$.

Using the Schur complement equivalence, (28) can be equivalent to be

$$\begin{bmatrix} \Gamma_5 & \Gamma_2 \\ \Gamma_2^T & \Gamma_3 \end{bmatrix} \prec 0, \tag{29}$$

where $\Gamma_5 = QE + E^T Q + k_1 I + Q\tilde{B}_\Phi \tilde{B}_\Phi^T Q + Q\tilde{B}_u \tilde{B}_u^T Q + QQ + \tilde{N}_{\Phi x}^T \tilde{N}_{\Phi x} + \tilde{N}_{vx}^T F_{\alpha\beta} \tilde{N}_{vx}$.

Split (29) into two matrices J_1 and J_2 such that

$$\begin{aligned} J_1 &= \begin{bmatrix} \Gamma_6 + \tilde{N}_{\Phi x}^T \tilde{N}_{\Phi x} & \tilde{N}_{\Phi x}^T \tilde{N}_{\Phi\omega} \\ \tilde{N}_{\Phi\omega}^T \tilde{N}_{\Phi x} & \tilde{N}_{\Phi\omega}^T \tilde{N}_{\Phi\omega} \end{bmatrix}, \\ J_2 &= \begin{bmatrix} \tilde{N}_{vx}^T F_{\alpha\beta} \tilde{N}_{vx} & \Gamma_7 \\ \tilde{N}_{v\omega}^T F_{\alpha\beta} \tilde{N}_{vx} + F_{\alpha+\beta} \tilde{N}_{vx} & \Gamma_8 \end{bmatrix}, \end{aligned} \tag{30}$$

where $\Gamma_6 = QE + E^T Q + k_1 I + Q\tilde{B}_\Phi \tilde{B}_\Phi^T Q + Q\tilde{B}_u \tilde{B}_u^T Q + QQ$, $\Gamma_7 = \tilde{N}_{vx}^T F_{\alpha\beta} \tilde{N}_{v\omega} + \tilde{N}_{vx}^T F_{\alpha+\beta}$, $\Gamma_8 = \tilde{N}_{v\omega}^T F_{\alpha\beta} \tilde{N}_{v\omega} + F_{\alpha+\beta} \tilde{N}_{v\omega} + \tilde{N}_{v\omega}^T F_{\alpha+\beta} + F_\lambda$.

Clearly, $J_1 + J_2 \prec 0$ and the following relation holds

$$\begin{aligned} J_1 &= \begin{bmatrix} I & 0 \\ \tilde{N}_{\Phi x} & \tilde{N}_{\Phi\omega} \end{bmatrix}^T \begin{bmatrix} \Gamma_6 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ \tilde{N}_{\Phi x} & \tilde{N}_{\Phi\omega} \end{bmatrix}, \\ J_2 &= \begin{bmatrix} \tilde{N}_{vx} & \tilde{N}_{v\omega} \\ 0 & I \end{bmatrix}^T \begin{bmatrix} F_{\alpha\beta} & F_{\alpha+\beta} \\ F_{\alpha+\beta} & F_\lambda \end{bmatrix} \begin{bmatrix} \tilde{N}_{vx} & \tilde{N}_{v\omega} \\ 0 & I \end{bmatrix}. \end{aligned} \tag{31}$$

From (19), multiplying of the matrix inequality $J_1 + J_2 \prec 0$ left and right by $[(x - \underline{x})^T, (\overline{x} - x)^T, (\omega_\Phi - \underline{\omega}_\Phi)^T, (\overline{\omega}_\Phi - \omega_\Phi)^T]$ and its transpose, we have

$$\begin{bmatrix} x - \underline{x} \\ \overline{x} - x \\ \Phi - \underline{\Phi} \\ \overline{\Phi} - \Phi \end{bmatrix}^T \begin{bmatrix} \Gamma_6 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x - \underline{x} \\ \overline{x} - x \\ \Phi - \underline{\Phi} \\ \overline{\Phi} - \Phi \end{bmatrix} + \Pi < 0. \tag{32}$$

Combining (25) in Theorem 3.1, we can conclude that

$$\begin{bmatrix} x - \underline{x} \\ \overline{x} - x \\ \Phi - \underline{\Phi} \\ \overline{\Phi} - \Phi \end{bmatrix}^T \begin{bmatrix} \Gamma_6 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x - \underline{x} \\ \overline{x} - x \\ \Phi - \underline{\Phi} \\ \overline{\Phi} - \Phi \end{bmatrix} < 0. \tag{33}$$

Then we consider the error dynamical system (9) and rewrite it as

$$\begin{aligned} \dot{\tilde{e}} &= \begin{bmatrix} A - \underline{L}C & 0 \\ 0 & A - \overline{L}C \end{bmatrix} \begin{bmatrix} \underline{e} \\ \overline{e} \end{bmatrix} + \begin{bmatrix} B_\Phi & 0 \\ 0 & B_\Phi \end{bmatrix} \begin{bmatrix} \Phi - \underline{\Phi} \\ \overline{\Phi} - \Phi \end{bmatrix} \\ &\quad + \begin{bmatrix} B_u & 0 \\ 0 & B_u \end{bmatrix} \begin{bmatrix} u - \underline{u} \\ \overline{u} - u \end{bmatrix} + \begin{bmatrix} f(x) - \underline{f}(\underline{x}, \overline{x}) \\ \overline{f}(\underline{x}, \overline{x}) - f(x) \end{bmatrix} \\ &= E\tilde{e} + \tilde{B}_\Phi \Delta\Phi + \tilde{B}_u \Delta u + \tilde{f}. \end{aligned} \tag{34}$$

To prove the stability of error dynamical system (9), let us consider a Lyapunov function $V(t) = \tilde{e}^T Q \tilde{e}$, whose time derivative takes the form:

$$\begin{aligned} \dot{V}(t) &= (\dot{\tilde{e}}^T Q\tilde{e} + \tilde{e}^T Q\dot{\tilde{e}}) \\ &= (E\tilde{e} + \tilde{B}_\Phi \Delta\Phi + \tilde{B}_u \Delta u + \tilde{f})^T Q\tilde{e} + \tilde{e}^T Q(E\tilde{e} + \tilde{B}_\Phi \Delta\Phi \\ &\quad + \tilde{B}_u \Delta u + \tilde{f}) \\ &= \tilde{e}^T (QE + E^T Q)\tilde{e} + 2\tilde{e}^T Q\tilde{B}_\Phi \Delta\Phi + 2\tilde{e}^T Q\tilde{B}_u \Delta u \\ &\quad + 2\tilde{f}^T Q\tilde{e}. \end{aligned} \tag{35}$$

The following inequalities are introduced

$$2\tilde{e}^T Q\tilde{B}_\Phi \Delta\Phi \le \tilde{e}^T Q\tilde{B}_\Phi \tilde{B}_\Phi^T Q\tilde{e} + \Delta\Phi^T \Delta\Phi,$$

$$2\tilde{e}^T Q\tilde{B}_u \Delta u \le \tilde{e}^T Q\tilde{B}_u \tilde{B}_u^T Q\tilde{e} + \Delta u^T \Delta u,$$

$$2\tilde{f}^T Q\tilde{e} \le \tilde{e}^T QQ\tilde{e} + \tilde{f}^T \tilde{f}.$$

Under Assumption 2.3, it implies that

$$\begin{aligned} \tilde{f}^T\tilde{f} &= (\underline{a}_1\underline{e} + \underline{a}_2\overline{e} + \underline{\rho})^T(\underline{a}_1\underline{e} + \underline{a}_2\overline{e} + \underline{\rho}) \\ &\quad + (\overline{a}_1\underline{e} + \overline{a}_2\overline{e} + \overline{\rho})^T(\overline{a}_1\underline{e} + \overline{a}_2\overline{e} + \overline{\rho}) \\ &= \|\underline{a}_1\underline{e} + \underline{a}_2\overline{e} + \underline{\rho}\|^2 + \|\overline{a}_1\underline{e} + \overline{a}_2\overline{e} + \overline{\rho}\|^2 \\ &\le 3(\underline{a}_1^2\|\underline{e}\|^2 + \underline{a}_2^2\|\overline{e}\|^2 + \|\underline{\rho}\|^2) \\ &\quad + 3(\overline{a}_1^2\|\underline{e}\|^2 + \overline{a}_2^2\|\overline{e}\|^2 + \|\overline{\rho}\|^2) \\ &= \begin{bmatrix} \underline{e} \\ \overline{e} \end{bmatrix}^T \begin{bmatrix} 3(\underline{a}_1^2 + \overline{a}_1^2) & 0 \\ 0 & 3(\underline{a}_2^2 + \overline{a}_2^2) \end{bmatrix} \begin{bmatrix} \underline{e} \\ \overline{e} \end{bmatrix} + 3(\|\underline{\rho}\|^2 + \|\overline{\rho}\|^2) \\ &\le \tilde{e}^T k_1 I\tilde{e} + 3k_2, \end{aligned}$$

where $k_1 = 3max\{(\underline{a}_1^2 + \overline{a}_1^2), (\underline{a}_2^2 + \overline{a}_2^2)\}, k_2 = \|\underline{\rho}\|^2 + \|\overline{\rho}\|^2$.

Therefore, (35) implies that

$$\begin{aligned} \dot{V}(t) &\le \tilde{e}^T \Gamma_7\tilde{e} + \Delta\Phi^T \Delta\Phi + \Delta u^T \Delta u + \tilde{f}^T\tilde{f} \\ &\le \tilde{e}^T \Gamma_7\tilde{e} + \Delta\Phi^T \Delta\Phi + \tilde{e}^T k_1 I\tilde{e} + \Delta u^T \Delta u + 3k_2 \\ &= \tilde{e}^T (\Gamma_7 + k_1 I)\tilde{e} + \Delta\Phi^T \Delta\Phi + \Delta u^T \Delta u + 3k_2 \\ &= \tilde{e}^T \Gamma_6\tilde{e} + \Delta\Phi^T \Delta\Phi + \Delta u^T \Delta u + 3k_2, \end{aligned}$$

where $\Gamma_7 = QE + E^T Q + Q\tilde{B}_\Phi\tilde{B}_\Phi^T Q + Q\tilde{B}_u\tilde{B}_u^T Q + QQ$.

According to (33), we can get the following condition

$$\begin{bmatrix} \tilde{e} \\ \Delta\Phi \end{bmatrix}^T \begin{bmatrix} \Gamma_6 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \tilde{e} \\ \Delta\Phi \end{bmatrix} < 0, \tag{36}$$

from which it can be inferred that

$$\tilde{e}^T \Gamma_6\tilde{e} + \Delta\Phi^T \Delta\Phi < 0. \tag{37}$$

According to the conditions (37) derived above, in any case, there must be a real number $\varepsilon > 0$, such that the following equation holds

$$\tilde{e}^T (\Gamma_6 + \varepsilon Q)\tilde{e} + \Delta\Phi^T \Delta\Phi < 0. \tag{38}$$

which can be rewritten to

$$\tilde{e}^T \Gamma_6\tilde{e} + \Delta\Phi^T \Delta\Phi < -\varepsilon\tilde{e}^T Q\tilde{e}. \tag{39}$$

Moreover, based on (36), we can conclude that

$$\dot{V}(t) \le -\varepsilon V + c_2, \tag{40}$$

where $c_2 \in \mathbb{R}_+$ and $c_2 \ge \Delta u^T \Delta u + 3k_2$. According to Lemma 2.3, the error dynamical system (9) is practically stable.

Adding or subtracting S does not affect the Metzler property of the expression because S is a diagonal matrix. Thus $Q\tilde{A} - M\tilde{C}$ is Metzler considering matrix inequality (27). Based on $M = Q\tilde{L}$, then $Q\tilde{A} - Q\tilde{L}\tilde{C}$ is Metzler. Multiplying the diagonal matrix Q will not change the Metzler features based on Lemma 2.1, so $\tilde{A} - \tilde{L}\tilde{C}$ is Metzler. Thus the proof is complete. ■

Remark 3.2: It is worth noting that considering the setting $\underline{x}(0) \le x(0) \le \overline{x}(0)$ mentioned in Assumption 2.2, it is possible the left side of (32) is equal to 0 when $t = 0$. Since $\underline{u} < u < \overline{u}$, according to the theory of positive systems, the state variable $\underline{e} = \overline{e} = 0$ in system (9) when and only when $t = 0$. This means that the case $\underline{x} = x = \overline{x}$ exists only when $t = 0$. Here we consider the fact that (32) does not hold only at this moment $t = 0$ and does not have an impact on the correctness of Theorem 3.2. The main reason why the case $\underline{x}(0) = x(0) = \overline{x}(0)$ is retained in the assumption on the initial value of the system state is to minimise the usage restrictions of the proposed method.

4. Application to lateral vehicle control systems

In this section, the developed run-time safety monitor design methodology is applied to the lateral vehicle control system to evaluate the correctness and applicability of the proposed methodology. The National Highway Transportation Safety Administration (NHTSA) has identified lane departures as the leading cause of rollovers in sport utility vehicles (SUVs) and light trucks (http://www.nhtsa.gov). Lateral vehicle control is an important approach to resolving lane departure accidents and has been heavily researched in industry and academia. Lateral vehicle control means that the vehicle collects road and environmental information via sensors such as magnetic materials, vision systems, or GPS to obtain the vehicle's position relative to the desired path. Control commands are then issued to the vehicle based on a control strategy. The control process can be summarised into two parts: detection and reaction. The detection device evaluates the position of the vehicle relative to the road in real time and determines whether a road deviation has occurred. Once a deviation is detected, the controller issues a warning to the driver and/or intervenes in the vehicle.

In the following example, we consider a 'bicycle' model of a vehicle with two degrees of freedom, the lateral position Y of the vehicle, and the yaw angle θ of the vehicle, as shown in Figure 2. This system is under the control of a neural network controller, which serves to provide intervention when the vehicle leaves the road centre line. Consider the lateral vehicle control system from Rajamani (2011):

$$\mathcal{L}_{eg} : \begin{cases} \dot{x} = Ax + B_\Phi \Phi(x) + B_u u \\ y = Cx \end{cases}, \tag{41}$$

where system matrices A, B_Φ, B_u, C, and input u are defined by

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -\frac{2C_{af}+2C_{ar}}{mV_x} & \frac{2C_{af}+2C_{ar}}{m} & \frac{-2C_{af}l_f+2C_{ar}l_r}{mV_x} \\ 0 & 0 & 0 & 1 \\ 0 & -\frac{2C_{af}l_f-2C_{ar}l_r}{I_zV_x} & \frac{2C_{af}l_f-2C_{ar}l_r}{I_z} & -\frac{2C_{af}l_f^2+2C_{ar}l_r^2}{I_zV_x} \end{bmatrix},$$

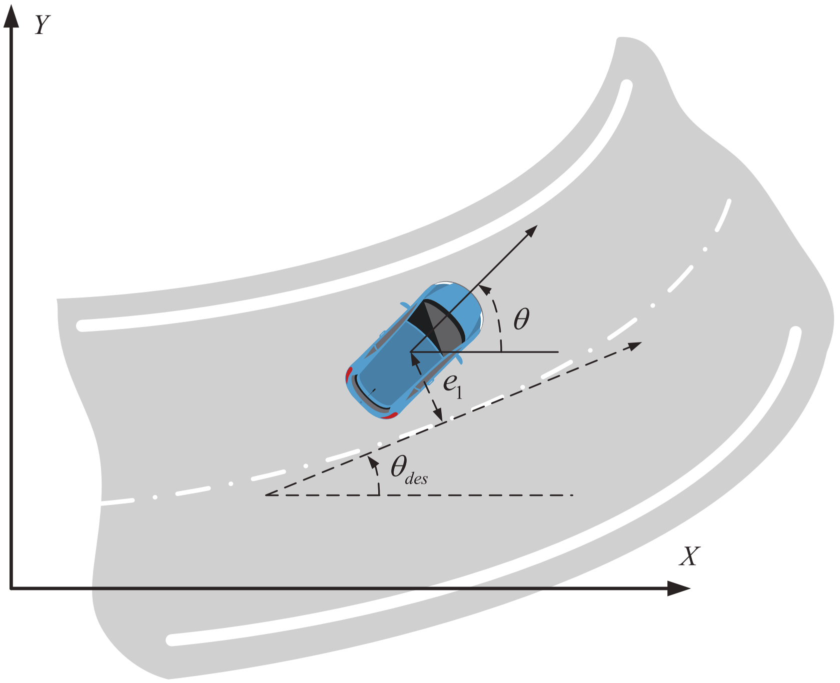


Figure 2. Illustration of Lateral Vehicle Control System.

$$B_\Phi = \begin{bmatrix} 0 \\ \frac{2C_{af}}{m} \\ 0 \\ \frac{2C_{af}l_f}{I_z} \end{bmatrix},$$

$$B_u = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$u = \begin{bmatrix} 0 \\ -\frac{2C_{af}l_f - 2C_{ar}l_r}{mV_x} - V_x \\ 0 \\ -\frac{2C_{af}l_f^2 + 2C_{ar}l_r^2}{I_z V_x} \end{bmatrix} \dot{\theta}_{des}.$$

Let $x = [e_1, \dot{e}_1, e_2, \dot{e}_2]^T$ denote the state of the system (41). e_1 is the distance of the c.g. (centre of gravity) of the vehicle from the centreline of the lane (m). e_2 is the orientation error of the vehicle with respect to the road (rad), which can be obtained by the equation $e_2 = \theta - \theta_{des}$. θ is called the heading angle of the vehicle, and θ_{des} is the desired orientation of the vehicle, with respect to the global X-axis (rad). $\dot{\theta}_{des} = \frac{V_x}{R}$ is defined as the rate of change of the desired orientation of the vehicle(rad/s). $\Phi(x)$ is the neural network controller, and its output is the front wheel steering angle (rad). System parameters are given in Table 2. By calculation to u, we set $\underline{u} = [-1, -3, -1, -1]^T, \overline{u} = [1, -1, 1, 0]^T$.

The lateral vehicle control system we are discussing does not contain nonlinear functions, which means $f(x) = \underline{f}(\underline{x}, \overline{x}) = \overline{f}(\underline{x}, \overline{x}) = 0$, so the corresponding interval observer system is as follows

$$\mathcal{M}_{eg} : \begin{cases} \dot{\underline{x}} = (A - \underline{L}C)\underline{x} + \underline{L}y + B_\Phi \underline{\Phi}(\underline{x}, \overline{x}) + \underline{u} \\ \dot{\overline{x}} = (A - \overline{L}C)\overline{x} + \overline{L}y + B_\Phi \overline{\Phi}(\underline{x}, \overline{x}) + \overline{u} \end{cases}.$$

Table 2. System Parameters for Lateral Vehicle Control System.

Total mass of vehicle	$m = 1573$ kg
Yaw moment of inertia of vehicle	$I_z = 2873$ kg · m^2
Longitudinal distance from c.g. to front tires	$l_f = 1.1$ m
Longitudinal distance from c.g. to rear tires	$l_r = 1.58$ m
Front tire cornering stiffness	$C_{af} = 80000$ N/rad
Rear tire cornering stiffness	$C_{ar} = 80000$ N/rad
Longitudinal velocity of the c.g. of the vehicle	$V_x = 30$ m/s
Constant road radius	$R = 400$ m

The following describes how the auxiliary neural networks, $\underline{\Phi}(\underline{x}, \overline{x})$ and $\overline{\Phi}(\underline{x}, \overline{x})$, and the interval observer gains, $\underline{L}$ and $\overline{L}$, are obtained in this example. According to the feedback gain K given in paper (Alleyne, 1997), the system (41) can operate normally. Based on this operating data, we train the neural network controller $\Phi(x)$, which is parameterised by a 3-layer feedforward neural network with $n_1 = 5$, $n_2 = 5$, and $n_3 = 1$, and $\psi(v) = tanh(v)$ as the activation function of the first two layers. The third layer does not use the activation function according to the settings in our paper. The auxiliary neural networks $\underline{\Phi}(\underline{x}, \overline{x})$ and $\overline{\Phi}(\underline{x}, \overline{x})$ are designed based on $\Phi(x)$ according to (14) and (15). Considering the physical limitations of vehicle dynamics, the range of front wheel steering angles is limited to $[-\pi/6, \pi/6]$, which means that the output of neural network $\Phi(x)$ and auxiliary neural networks, $\underline{\Phi}(\underline{x}, \overline{x})$ and $\overline{\Phi}(\underline{x}, \overline{x})$, are limited to $[-\pi/6, \pi/6]$. The observer gains, $\underline{L}$ and $\overline{L}$, can be obtained by solving linear matrix inequalities (26) and (27) in Theorem 3.2.

The run-time boundary estimations of state trajectories of lateral position error $\{e_1, \dot{e}_1\}$ and yaw angle error $\{e_2, \dot{e}_2\}$ during the lateral vehicle control system evolves in time interval $[0, 10]$ are shown in Figures 3 and 4. The lateral position error and yaw angle error decrease significantly after the system reaches a steady state, indicating that the original system operates normally under the action of the neural network controller $\Phi(x)$.

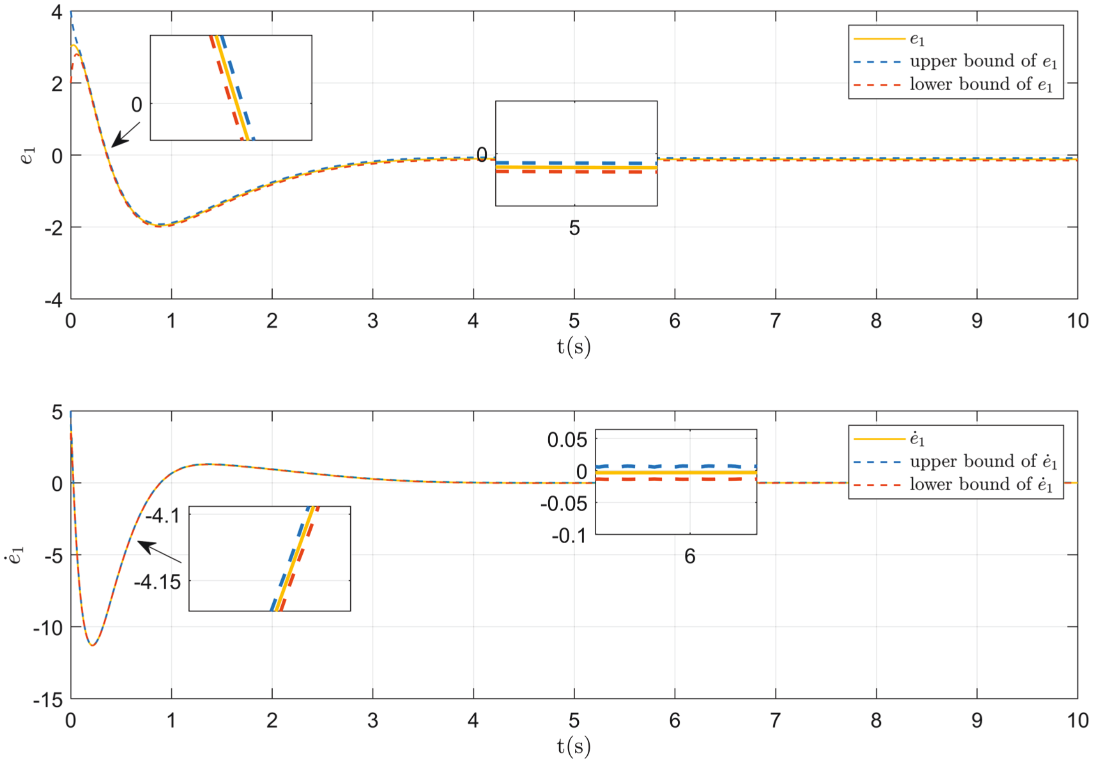


Figure 3. Safety monitoring of lateral position error e_1 and its derivative $\dot{e}_1$.

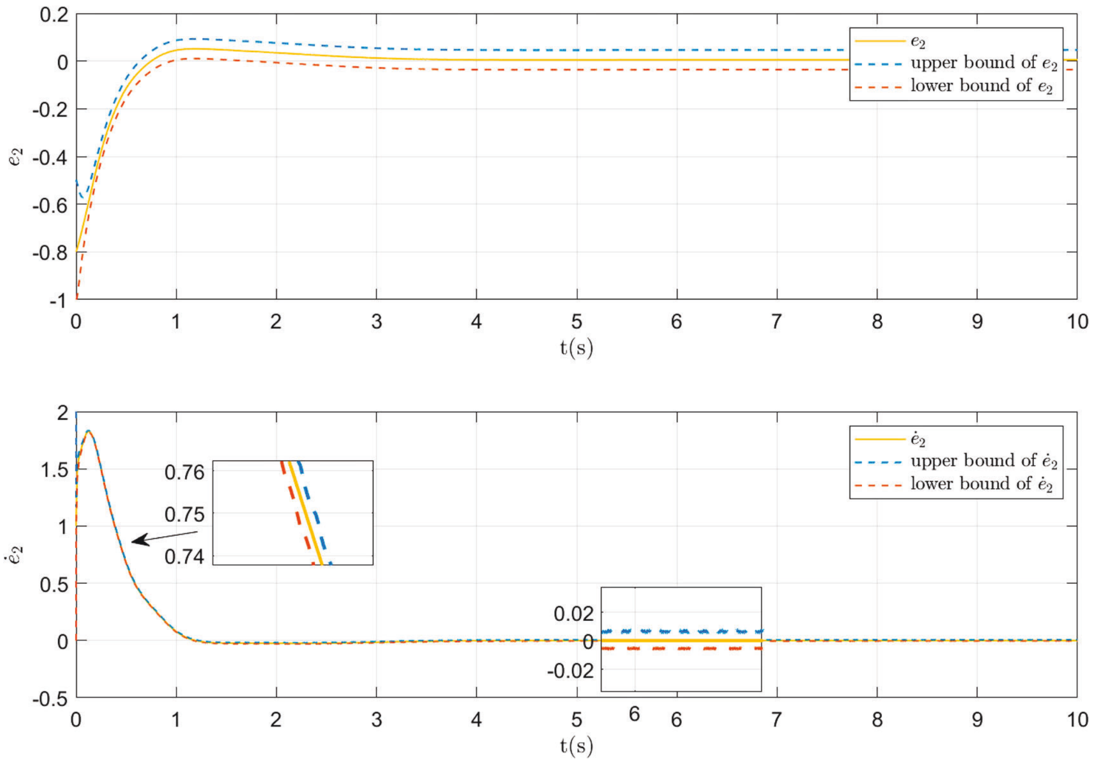


Figure 4. Safety monitoring of yaw angle error e_2 and its derivative $\dot{e}_2$.

It is worth noting that the steady-state values of e_1 and e_2 are not zero because the input due to road curvature $\dot{\theta}_{des}$ is non-zero. The specific physical explanation of these steady-state errors can be found in Sections 3.2 and 3.3 of Rajamani (2011). As shown in the results, the state trajectories (solid line) always run between the upper and lower bounds of the interval observer (dashed line), indicating that the interval observer we have designed can be used for state safety monitoring.

5. Conclusions

This paper presents a possible solution to the problem of run-time safety monitoring of dynamical systems embedded with neural network components. A design approach for a safety monitor is proposed for the system characteristics. The safety monitor works as a Luenberger-type interval observer, which estimates the upper and lower bounds of the state run-time trajectory in real time. The design process of the interval observer consists of two main components: the two auxiliary neural networks and the observer gain. The two auxiliary neural networks can be obtained from the neural network embedded in the original system. The presence of nonlinear activation functions in neural networks makes it difficult to apply traditional control theory to calculate observer gains $\underline{L}$ and $\overline{L}$. To solve this problem, we use quadratic constraints (QCs) to abstract the nonlinear activation functions in neural networks. The computational problem of observer gain is expressed in a series of convex optimisation problems. The interval observer design method is applied to the lateral vehicle control system to verify the correctness of the proposed solutions. The correction of neural network operation in the event of security problems needs to be considered in future work. Further applications to dynamical systems with more complex behaviours such as switched or hybrid systems (Li et al., 2020; Xiang et al., 2017a; Zhu et al., 2019) will be also considered in the future.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Science Foundation, under NSF CAREER Award 2143351, NSF CNS Award no. 2223035, and NSF IIS Award no. 2331938.

References

Alleyne, A. (1997). A comparison of alternative intervention strategies for unintended roadway departure (URD) control. *Vehicle System Dynamics*, *27*(3), 157–186. https://doi.org/10.1080/00423119708969327

Anderson, C. W., Young, P. M., Buehner, M. R., Knight, J. N., Bush, K. A., & Hittle, D. C. (2007). Robust reinforcement learning control using integral quadratic constraints for recurrent neural networks. *IEEE Transactions on Neural Networks*, *18*(4), 993–1002. https://doi.org/10.1109/TNN.2007.899520

Bolajraf, M., Rami, M. A., & Helmke, U. (2011). Robust positive interval observers for uncertain positive systems. *IFAC Proceedings Volumes*, *44*(1), 14330–14334. https://doi.org/10.3182/20110828-6-IT-1002.03682

Cacace, F., Germani, A., & Manes, C. (2015). A new approach to design interval observers for linear systems. *IEEE Transactions on Automatic Control*, *60*(6), 1665–1670. https://doi.org/10.1109/TAC.2014.2359714

Chebotarev, S., Efimov, D., Raïssi, T., & Zolghadri, A. (2015). Interval observers for continuous-time LPV systems with L_1/L_2 performance. *Automatica*, *58*, 82–89. https://doi.org/10.1016/j.automatica.2015.05.009

Dutta, S., Chen, X., & Sankaranarayanan, S. (2019). Reachability analysis for neural feedback systems using regressive polynomial rule inference. In *Proceedings of the 22nd ACM international conference on hybrid systems: Computation and control* (pp. 157–168).

Dutta, S., Jha, S., Sankaranarayanan, S., & Tiwari, A. (2018, April 17–19). Output range analysis for deep feedforward neural networks. In Aaron Dutle, César Muñoz, and Anthony Narkawicz (Eds.), *Nasa formal methods symposium* (pp. 121–138). Newport News, VA, USA.

Efimov, D., & Raïssi, T. (2016). Design of interval observers for uncertain dynamical systems. *Automation and Remote Control*, *77*(2), 191–225. https://doi.org/10.1134/S0005117916020016

Fazlyab, M., Morari, M., & Pappas, G. J. (2022). Safety verification and robustness analysis of neural networks via quadratic constraints and semidefinite programming. *IEEE Transactions on Automatic Control*, *67*(1), 1–15. https://doi.org/10.1109/TAC.2020.3046193

Fazlyab, M., Robey, A., Hassani, H., Morari, M., & Pappas, G. (2019). Efficient and accurate estimation of lipschitz constants for deep neural networks. *Proceedings of Advances Neural Information Processing Systems*, *2019*, 11423–11434.

Ge, S., & Wang, C. (2004). Adaptive neural control of uncertain MIMO nonlinear systems. *IEEE Transactions on Neural Networks*, *15*(3), 674–692. https://doi.org/10.1109/TNN.2004.826130

Hu, H., Fazlyab, M., Morari, M., & Pappas, G. J. (2020). Reach-SDP: Reachability analysis of closed-loop systems with neural network controllers via semidefinite programming. In *2020 59th IEEE conference on decision and control (CDC)* (pp. 5929–5934). https://doi.org/10.1109/CDC42340.2020.9304296

Hunt, K., Sbarbaro, D., Żbikowski, R., & Gawthrop, P. (1992). Neural networks for control systems—A survey. *Automatica*, *28*(6), 1083–1112. https://doi.org/10.1016/0005-1098(92)90053-I

Levin, A., & Narendra, K. (1992). Stabilization of nonlinear dynamical systems using neural networks. In *International joint conference on neural networks* (Vol. 1, pp. 275–280). https://doi.org/10.1109/IJCNN.1992.287122

Li, S., Ahn, C. K., Guo, J., & Xiang, Z. (2020). Neural-network approximation-based adaptive periodic event-triggered output-feedback control of switched nonlinear systems. *IEEE Transactions on Cybernetics*, *51*(8), 4011–4020. https://doi.org/10.1109/TCYB.2020.3022270

Lomuscio, A., & Maganti, L. (2017). *An approach to reachability analysis for feed-forward ReLU neural networks*. arXiv preprint arXiv:1706.07351.

Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal adversarial perturbations. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 86–94). https://doi.org/10.1109/CVPR.2017.17

Niu, B., Liu, Y., Zhou, W., Li, H., Duan, P., & Li, J. (2020). Multiple lyapunov functions for adaptive neural tracking control of switched nonlinear nonlower-triangular systems. *IEEE Transactions on Cybernetics*, *50*(5), 1877–1886. https://doi.org/10.1109/TCYB.6221036

Rajamani, R. (2011). *Vehicle dynamics and control*. Springer Science & Business Media.

Takahashi, K. (2017). Remarks on adaptive-type hypercomplex-valued neural network-based feedforward feedback controller. In *2017 IEEE international conference on computer and information technology (CIT)* (pp. 151–156). https://doi.org/10.1109/CIT.2017.16

Tran, H. D., Manzanas Lopez, D., Musau, P., Yang, X., Nguyen, L. V., Xiang, W., & Johnson, T. T. (2019). Star-based reachability analysis of deep neural networks. In *International symposium on formal methods* (pp. 670–686).

Tran, H. D., Musau, P., Manzanas Lopez, D., Yang, X., Nguyen, L. V., Xiang, W., & Johnson, T. T. (2019). Parallelizable reachability analysis algorithms for feed-forward neural networks. In *2019 IEEE/ACM 7th international conference on formal methods in software engineering (formalise)* (pp. 51–60). https://doi.org/10.1109/FormaliSE.2019.00012

Wang, T., Li, Y., & Xiang, W. (2022). Design of interval observer for continuous linear large-scale systems with disturbance attenuation. *Journal of the Franklin Institute*, *359*(8), 3910–3929.

Wu, Z. G., Shi, P., Su, H., & Chu, J. (2014). Exponential stabilization for sampled-data neural-network-based control systems. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(12), 2180–2190. https://doi.org/10.1109/TNNLS.2014.2306202

Xiang, W. (2021). Runtime safety monitoring of neural-network-enabled dynamical systems. *IEEE Transactions on Cybernetics*, 1–10. https://doi.org/10.1109/TCYB.2021.3053575

Xiang, W., Tran, H. D., & Johnson, T. T. (2017a). Output reachable set estimation for switched linear systems and its application in safety verification. *IEEE Transactions on Automatic Control*, *62*(10), 5380–5387. https://doi.org/10.1109/TAC.2017.2692100

Xiang, W., Tran, H. D., & Johnson, T. T. (2017b). *Reachable set computation and safety verification for neural networks with ReLU activations*. arXiv preprint arXiv:1712.08163.

Xiang, W., Tran, H. D., & Johnson, T. T. (2018). Output reachable set estimation and verification for multilayer neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(11), 5777–5783. https://doi.org/10.1109/TNNLS.2018.2808470

Xiang, W., Tran, H. D., Rosenfeld, J. A., & Johnson, T. T. (2018). Reachable set estimation and safety verification for piecewise linear systems with neural network controllers. In *2018 annual American control conference (ACC)* (pp. 1574–1579). https://doi.org/10.23919/ACC.2018.8431048

Xiang, W., Tran, H. D., Yang, X., & Johnson, T. T. (2021). Reachable set estimation for neural network control systems: A simulation-guided approach. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(5), 1821–1830. https://doi.org/10.1109/TNNLS.2020.2991090

Yin, H., Seiler, P., & Arcak, M. (2022). Stability analysis using quadratic constraints for systems with neural network controllers. *IEEE Transactions on Automatic Control*, *67*(4), 1980–1987. https://doi.org/10.1109/TAC.2021.3069388

Zhang, D., Couto, L. D., Gill, P., Benjamin, S., Zeng, W., & Moura, S. J. (2020). Interval observer for SOC estimation in parallel-connected lithium-ion batteries. In *2020 American control conference (ACC)* (pp. 1149–1154). https://doi.org/10.23919/ACC45564.2020.9147468

Zhang, L., Zhu, Y., & W. X. Zheng (2017). State estimation of discrete-time switched neural networks with multiple communication channels. *IEEE Transactions on Cybernetics*, *47*(4), 1028–1040. https://doi.org/10.1109/TCYB.2016.2536748

Zhang, Yw., Shen, X., Xu, Hf., Lin, Zr., Ni, Hy., & Yan, Wx. (2020). Fault diagnosis of multi-area automatic generation control via interval observer technique. In *2020 IEEE international conference on high voltage engineering and application (ICHVE)* (pp. 1–4). https://doi.org/10.1109/ICHVE49031.2020.9279463

Zheng, G., Efimov, D., & Perruquetti, W. (2016). Design of interval observer for a class of uncertain unobservable nonlinear systems. *Automatica*, *63*, 167–174. https://doi.org/10.1016/j.automatica.2015.10.007

Zhu, Y., Zheng, W. X., & Zhou, D. (2019). Quasi-synchronization of discrete-time Lur'e-type switched systems with parameter mismatches and relaxed PDT constraints. *IEEE Transactions on Cybernetics*, *50*(5), 2026–2037. https://doi.org/10.1109/TCYB.6221036