Guaranteed Neural-Network-based control for dynamical systems

Louis JOURET

Supervised by
Pr. Adnane Saoud (CentraleSupélec)
Pr. Aude Billard (EPFL)

August 2023

Contents

1 Introduction

In recent years, the development of deep learning techniques has led to significant advancements in the field of control systems. One such advancement is the use of neural networks as controllers, which have been widely used in various applications due to their ability to learn complex non-linear control policies for high-dimensional state and action spaces alwhile being computationaly efficient in the foward pass. However, the safety of neural network controllers has become a growing concern, as their behavior can be unpredictable and difficult to interpret.
In this master thesis, we aim to apply more traditional and theoretical methods of control systems to investigate and establish the safety of neural network controllers. Specifically, we will analyze the stability and robustness properties of neural network controllers using rigorous mathematical methods, such as Lyapunov theory and viability theory. By using viability theory, we will investigate the regions of state space in which the neural network controller guarantees safe operation, taking into account constraints on the state and input variables.
Our study is motivated by the increasing demand for autonomous systems in various domains, such as robotics, autonomous driving, and aerospace. The safety of these systems is of paramount importance, and our work will contribute to ensuring their safe operation by providing a rigorous analysis of neural network controllers.

$$\dot{x} = Ax + Bu \tag{1}$$
$$x^+ = (I + dt \cdot A)x + dt \cdot Bu \tag{2}$$

$$a(x) = \max(0.001x, x) \tag{3}$$

$$a(x) = \begin{cases} 0.001x - 0.999, & \text{if } x < -1.5 \\ 0.5x - 0.25, & \text{if } -1.5 < x < -0.5 \\ x, & \text{if } -0.5 < x < 0.5 \\ 0.5x + 0.25, & \text{if } 0.5 > x > 1.5 \\ 0.001x + 0.999, & \text{if } x > 1.5 \end{cases} \tag{4}$$

$$\text{Loss} = -Q(s, a) \tag{5}$$

$$\text{Loss} = \text{MSE}\{Q(s, a) - (r + \gamma Q(s^+, a^+))\} \tag{6}$$

$$u(\vec{x}) = NN_{actor}(\vec{x}) \tag{7}$$
$$= W_{a(\vec{x})} \cdot \vec{x} + \vec{b}_{a(\vec{x})} \tag{8}$$

$$\dot{\vec{x}} = A\vec{x} + B\vec{u} \tag{9}$$
$$= A\vec{x} + B(W_{a(\vec{x})} \cdot \vec{x} + \vec{b}_{a(\vec{x})}) \tag{10}$$
$$= (A + BW_{a(\vec{x})})\vec{x} + B\vec{b}_{a(\vec{x})} \tag{11}$$

2 Theory

Theorem 1 *Consider the closed set $\mathcal{O}$ and the system $\dot{x}(t) = f(x(t))$ and assume that for each initial condition $x(0) \notin \mathcal{O}$ it admits a unique solution defined for all $t \geq 0$. Then $x(t) \notin \mathcal{O}^\circ$ for $t \geq 0$ if and only if the velocity vector satisfies:*

$$f(x) \notin \mathcal{T}_{\mathcal{O}}(x)^\circ, \textit{for all } x \in \partial\mathcal{O}$$

Theorem 2 *Consider the pratical set $\mathcal{O}$ defined by*

$$\mathcal{O} = \left\{x : g_k(x) \leq 0, k = 1, 2, \ldots, r\right\} \tag{12}$$

Now consider the system $\dot{x}(t) = f(x(t))$ and assume that for each initial condition $x(0) \notin \mathcal{O}$ it admits a unique solution defined for all $t \geq 0$. Then $x(t) \notin \mathcal{O}^\circ$ for $t \geq 0$ if and only if the velocity vector satisfies:

$$f(x) \in \left\{z : \nabla g_i(x)^T z \geq 0, \textit{ for all } i \in B(x)\right\}, \textit{where } B(x) = \left\{i : g_i(x) = 0\right\}$$

Corollary 2.1 *If all the assumptions and conditions of Theorem 1 are verified and the set $\mathcal{O}$ is a polytope defined by*

$$\mathcal{O} = \{x : C_{\mathcal{O}} x \leq d_{\mathcal{O}}\} \tag{13}$$

then $x(t) \notin \mathcal{O}^\circ$ for $t \geq 0$ if and only if the velocity vector satisfies:

$$f(x) \in \left\{ z : \begin{bmatrix} \mathcal{C}_{\mathcal{O}_{i,0}} \\ . \\ . \\ . \\ \mathcal{C}_{\mathcal{O}_{i,n}} \end{bmatrix}^T \quad z > 0, \textit{ for all } i : row_i(\mathcal{C}_{\mathcal{O}}) \cdot x = \vec{d}_{\mathcal{O}_i} \right\} \tag{14}$$

Theorem 3 *Consider a linear time-invariant system of the form $\dot{\vec{x}} = Ax + \vec{b}$. Let's define $\mathcal{O} = \left\{x : c^T x \leq d\right\}$ and $x_\lambda = \lambda \cdot x_1 + (1 - \lambda) \cdot x_2$ where λ is a scalar.*
Then,

$$\begin{cases} c^T \cdot f(x_1) > 0 \\ c^T \cdot f(x_2) > 0 \end{cases} \Rightarrow c^T \cdot f(x_\lambda) > 0 \tag{15}$$