Reachability Analysis of Neural Network Controlled Systems: A Mixed Monotone Contracting Approach

Saber Jafarpour

Decision and Control Laboratory
Georgia Institute of Technology

June 23, 2023

Acknowledgment

Akash Harapanahalli
Georgia Tech

Alexander Davydov
UCSB

Francesco Bullo
UCSB

Samuel Coogan
Georgia Tech

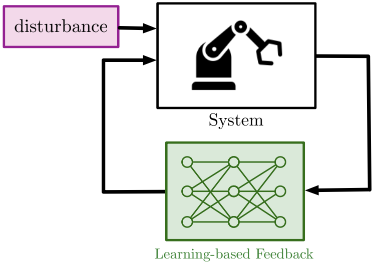
SJ and A. Harapanahalli and S. Coogan. Interval Reachability of Nonlinear Dynamical Systems with Neural Network Controllers. L4DC, 2023.

SJ and A. Davydov and F. Bullo. Non-Euclidean Contraction Theory for Monotone and Positive Systems. IEEE Transactions on Automatic Control, Dec. 2022

SJ and S. Coogan. Monotonicity and contraction on polyhedral cones. arXiv, 2021.

- Neural Networks as controllers in safety-critical applications (self driving vehicle and mobile robots)

Goal: ensure and verify *safety* of the closed-loop system



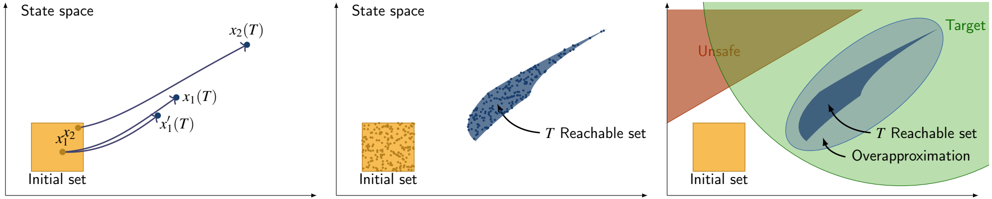
Features of neural network controllers:

- large # of parameters with nonlinearity
- sensitive wrt to input perturbations
- limited closed-loop safety guarantees

Challenges

Sound and computationally efficient methods for safety verification

System : $\dot{x} = f(x, w)$ **State** : $x \in \mathbb{R}^n$ **Disturbance** : $w \in \mathcal{W} \subseteq \mathbb{R}^m$



- reachable sets characterize evolution of the system

$$\mathcal{R}^f(t, \mathcal{X}_0) = \{x_w(t) \mid x_w(\cdot) \text{ is a traj of the system for some } w \text{ with } x_0 \in \mathcal{X}_0\}$$

- over-approximation of reachable sets for safety and verification
- reachability of dynamical system is an old problem with several classical approaches

not scalable to large-scale nonlinear systems

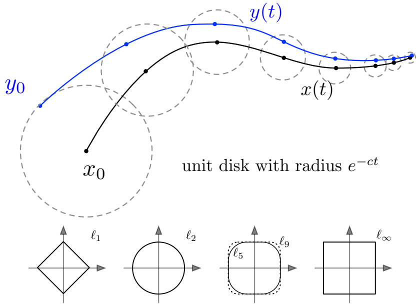
Contraction Theory

A framework for stability analysis

$\dot{x} = f(x, w)$ is contracting wrt $\|\cdot\|$ with rate c if
the dist between every two traj is decreasing/increasing with exp rate c wrt $\|\cdot\|$

Common applications

- convergence to reference trajectories
- efficient equilibrium point computation
- input-output robustness
- entrainment to periodic orbits



In this talk: contraction theory for reachability analysis

How to characterize contractivity using vector fields?

Matrix measure

Given a matrix $A \in \mathbb{R}^{n\times n}$ and a norm $\|\cdot\|$:

$$\mu_{\|\cdot\|}(A) := \lim_{h\to 0^+} \frac{\|I_n + hA\| - 1}{h}$$

Given $\eta \in \mathbb{R}^n_{\geq 0}$

$$\mu_{2,\eta}(A) = \tfrac{1}{2}\lambda_{\max}(\mathrm{diag}(\eta)A + A^\top \mathrm{diag}(\eta))$$

$$\mu_{1,\eta}(A) = \max_j \left(a_{jj} + \sum\nolimits_{i\neq j} |a_{ij}|\tfrac{\eta_j}{\eta_i}\right)$$

$$\mu_{\infty,\eta}(A) = \max_i \left(a_{ii} + \sum\nolimits_{j\neq i} |a_{ij}|\tfrac{\eta_j}{\eta_i}\right)$$

- directional derivative of matrix norm $\|\cdot\|$ in direction of A at point I_n,
- **In the literature**: one-sided Lipschitz constant, logarithmic norm

Classical result

$\dot{x} = f(x,w)$ is contracting wrt $\|\cdot\|$ with rate c iff

$$\mu_{\|\cdot\|}\left(\tfrac{\partial f}{\partial x}(x,w)\right) \leq c, \qquad \text{for all } x, w$$

Contraction-based Reachability Analysis

A global bound

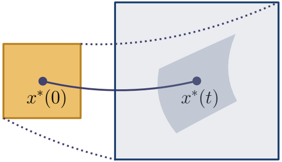
Assume $\mu_{\|\cdot\|}\left(\frac{\partial f}{\partial x}(x,w)\right) \leq c$ and $\left\|\frac{\partial f}{\partial w}(x,w)\right\| \leq \ell$

Theorem

If $\mathcal{X}_0 = B_{\|\cdot\|}(r_1, x_0^*)$ and $\mathcal{W} = B_{\|\cdot\|}(r_2, w^*)$, then

$$\mathcal{R}^f(t,\mathcal{X}_0) \subseteq B_{\|\cdot\|}(e^{ct}r_1 + \tfrac{\ell}{c}(e^{ct}-1)r_2, x^*(t))$$

where $x^*(\cdot)$ is the solution of $\dot{x} = f(x,w^*)$ with $x(0) = x_0^*$.



Proof: let $x(\cdot)$ be a traj of $\dot{x} = f(x,w)$. Using Taylor expansion, for $h \geq 0$

$$x(t+h) - x^*(t+h) = x(t) - x^*(t) + h\overbrace{\left(\int_0^1 D_x f(\tau x + (1-\tau)x^*)d\tau\right)}^{A(x,w)}(x(t)-x^*(t))$$

$$+ h\overbrace{\left(\int_0^1 D_w f(x, \tau w + (1-\tau)w^*)d\tau\right)}^{B(x,w)}(w-w^*) + \mathcal{O}(h^2)$$

Contraction-based Reachability

Proof continued

$$D^+\|x(t) - x^*(t)\| = \limsup_{h\to 0^+} \frac{\|x(t+h) - x^*(t+h)\| - \|x(t) - x^*(t)\|}{h}$$

$$= \limsup_{h\to 0^+} \frac{\|\left(I_n + hA(x,w)\right)\left(x(t) - x^*(t)\right) + hB(x,w)(w - w^*)\| - \|x(t) - x^*(t)\|}{h}$$

$$\leq \limsup_{h\to 0^+} \frac{\|\left(I_n + hA(x,w)\right)\left(x(t) - x^*(t)\right)\| + h\|B(x,w)\|\|w - w^*\| - \|x(t) - x^*(t)\|}{h}$$

$$\leq \mu_{\|\cdot\|}(A(x,w))\|x(t) - x^*(t)\| + \|B(x,w)\|\|w - w^*\|$$

$$\leq c\|x(t) - x^*(t)\| + \ell\|w - w^*\|$$

- generalized version of Grönwall's lemma
- overly conservative since c and ℓ are defined globally

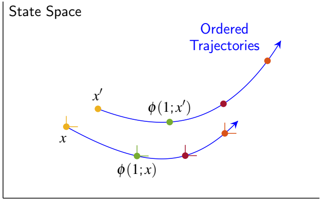
A dynamical system $\dot{x} = f(x, w)$ is monotone[1] if

$$x_u(0) \preceq_K y_w(0) \quad \text{and} \quad u \preceq_C w \quad \implies \quad x_u(t) \preceq_K y_w(t) \quad \text{for all time}$$

where $\preceq_K$ is the partial order with induced by the cone K.

A **polyhedral cone** has the form

$$K = \underbrace{\{y \in \mathbb{R}^n \mid H_K y \geq \mathbb{0}_p\}}_{\text{halfspace rep}} = \underbrace{\{V_K y \mid y \geq \mathbb{0}_p\}}_{\text{vertex rep}}$$



Monotonicity test

1. $H_K(\frac{\partial f}{\partial x}(x, w) + \alpha(x, w) I_n) V_K \geq \mathbb{0}_p$ for some $\alpha(x, w)$
2. $H_C \frac{\partial f}{\partial w}(x, w) V_C \geq \mathbb{0}_q$

[1] D. Angeli and E. Sontag, "Monotone control systems", TAC, 2003

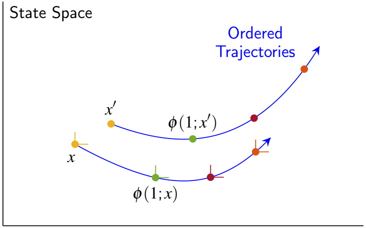
A dynamical system $\dot{x} = f(x, w)$ is monotone[1] if

$$x_u(0) \preceq_K y_w(0) \quad \text{and} \quad u \preceq_C w \quad \Longrightarrow \quad x_u(t) \preceq_K y_w(t) \quad \text{for all time}$$

where $\preceq_K$ is the partial order with induced by the cone K.

Monotonicity test (for standard partial order $\leq$)

1. $\frac{\partial f}{\partial x}(x, w)$ is Metzler (off-diag ≥ 0)
2. $\frac{\partial f}{\partial w}(x, w) \geq \mathbb{0}_m$



[1] D. Angeli and E. Sontag, "Monotone control systems", TAC, 2003

Theorem

For a monotone system with $\mathcal{W} = [\underline{w}, \overline{w}]$

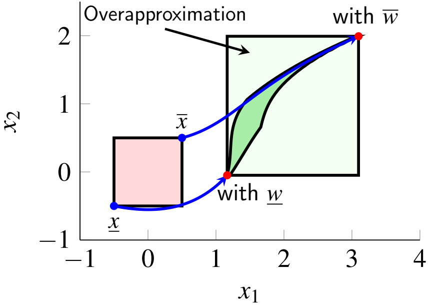
$$\mathcal{R}^f(t, [\underline{x}_0, \overline{x}_0]) \subseteq [x_{\underline{w}}(t), x_{\overline{w}}(t)]$$

where $x_{\underline{w}}(\cdot)$ (resp. $x_{\overline{w}}(\cdot)$) is the trajectory with disturbance $\underline{w}(\cdot)$ (resp. $\overline{w}(\cdot)$) starting at $\underline{x}_0$ (resp. $\overline{x}_0$)

Example:

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2^3 - x_1 + w \\ x_1 \end{bmatrix}$$

$$\mathcal{W} = [2.2\ ,\ 2.3] \quad \mathcal{X}_0 = \left[\begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix}, \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}\right]$$



Theorem

Let $\mathcal{X} = [\underline{x}_0, \overline{x}_0]$, $\mathcal{W} = [\underline{w}, \overline{w}]$, and $\eta \in \mathbb{R}^n_{\geq 0}$. Suppose that the monotone system $\dot{x} = f(x, w)$ is contracting with rate c wrt to $\|\cdot\|_{\infty,\eta}$, then

$$[x_{\underline{w}}(t), x_{\overline{w}}(t)] \subseteq B_{\|\cdot\|_{\infty,\eta}}(e^{ct} r_1 + \tfrac{\ell(e^{ct}-1)}{c} r_2, x^*(t))$$

where $x^*(\cdot)$ is the solution to $\dot{x} = f(x, w^*)$ with $x^*(0) = x_0^* \in \mathcal{X}_0$ and

$$r_1 = \max\{\|\underline{x} - x_0^*\|_{\infty,\eta}, \|\overline{x} - x_0^*\|_{\infty,\eta}\},$$
$$r_2 = \max\{\|\underline{w} - w^*\|_{\infty,\eta}, \|\overline{w} - w^*\|_{\infty,\eta}\}$$

Proof: both $x_{\underline{w}}(\cdot)$ and $x_{\overline{w}}(\cdot)$ are trajectories of the system

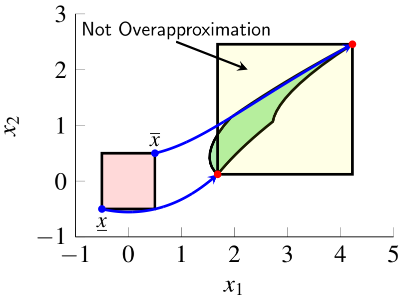
monotone reachability is at least as accurate as contraction reachability wrt diagonally weighted ℓ_∞-norms

- For non-monotone dynamical systems the extreme trajectories do not provide any over-approximation of reachable sets

Example:

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_2^3 - x_2 + w \\ x_1 \end{bmatrix}$$

$$\mathcal{W} = [2.2\ ,\ 2.3] \quad \mathcal{X}_0 = \left[\begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix}, \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}\right]$$



- **Key idea:** embed the dynamical system on $\mathbb{R}^n$ into a dynamical system on $\mathbb{R}^{2n}$
- Assume $\mathcal{W} = [\underline{w}, \overline{w}]$ and $\mathcal{X}_0 = [\underline{x}_0, \overline{x}_0]$

Original system

$$\dot{x} = f(x, w)$$

Embedding system

$$\dot{\underline{x}} = \underline{d}(\underline{x}, \overline{x}, \underline{w}, \overline{w}),$$
$$\dot{\overline{x}} = \overline{d}(\underline{x}, \overline{x}, \underline{w}, \overline{w})$$

$\underline{d}, \overline{d}$ are **decomposition functions** s.t.

1. $f(x, w) = \underline{d}(x, x, w, w)$ for every x, w
2. cooperative: $(\underline{x}, \underline{w}) \mapsto \underline{d}(\underline{x}, \overline{x}, \underline{w}, \overline{w})$
3. competitive: $(\overline{x}, \overline{w}) \mapsto \underline{d}(\underline{x}, \overline{x}, \underline{w}, \overline{w})$
4. the same properties for $\overline{d}$

The embedding system is a monotone dynamical system on $\mathbb{R}^{2n}$ with respect to the **southeast** partial order $\leq_{\mathrm{SE}}$:

$$\begin{bmatrix} x \\ \widehat{x} \end{bmatrix} \leq_{\mathrm{SE}} \begin{bmatrix} y \\ \widehat{y} \end{bmatrix} \quad \Longleftrightarrow \quad x \leq y \quad \text{and} \quad \widehat{y} \leq \widehat{x}$$

In terms of cones, $\leq_{\mathrm{SE}}$ is induced by the cone $\mathbb{R}^n_{\geq 0} \times -\mathbb{R}^n_{\geq 0}$.

- f locally Lipschitz $\implies$ a decomposition function exists
- decomposition functions are not unique (use structure of the system to construct one)

History:

J-L. Gouze and L. P. Hadeler. Monotone flows and order intervals. Nonlinear World, 1994

G. Enciso, H. Smith, and E. Sontag. Nonmonotone systems decom- posable into monotone systems with negative feedback . Journal of Differential Equations, 2006.

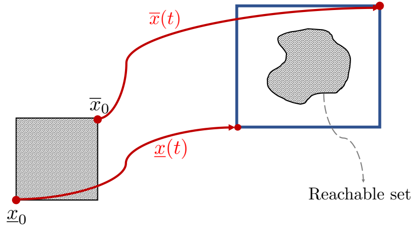
H. Smith. Global stability for mixed monotone systems. Journal of Difference Equations and Applications, 2008

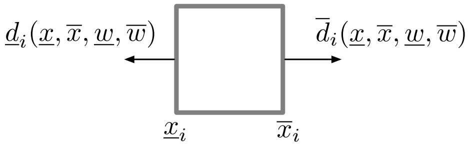
Theorem

Assume $\mathcal{W} = [\underline{w}_0, \overline{w}_0]$ and $\mathcal{X}_0 = [\underline{x}_0, \overline{x}_0]$ and

$$\dot{\underline{x}} = \underline{d}(\underline{x}, \overline{x}, \underline{w}, \overline{w}), \qquad \underline{x}(0) = \underline{x}_0$$
$$\dot{\overline{x}} = \overline{d}(\overline{x}, \underline{x}, \overline{w}, \underline{w}), \qquad \overline{x}(0) = \overline{x}_0$$

Then $\mathcal{R}^f(t, \mathcal{X}_0) \subseteq [\underline{x}(t), \overline{x}(t)]$





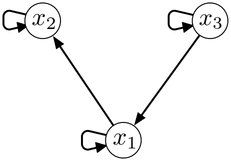
(Scalable) a single trajectory of embedding system provides **lower bound** ($\underline{x}$) and **upper bound** ($\overline{x}$) for the trajectories of the original system.

Embedding System for Linear Dynamical System

A structure preserving decomposition

- **Metzler/non-Metzler** decomposition: $A = \lceil A \rceil^{\mathrm{Mzl}} + \lfloor A \rfloor^{\mathrm{Mzl}}$
- Example: $A = \begin{bmatrix} 2 & 0 & -1 \\ 1 & -3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \implies \lceil A \rceil^{\mathrm{Mzl}} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & -3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \lfloor A \rfloor^{\mathrm{Mzl}} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

Linear systems

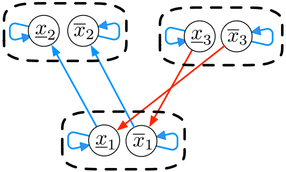


Original system

$$\dot{x} = Ax + Bw$$

Embedding system

$$\dot{\underline{x}} = \lceil A \rceil^{\mathrm{Mzl}} \underline{x} + \lfloor A \rfloor^{\mathrm{Mzl}} \overline{x} + B^{+} \underline{w} + B^{-} \overline{w}$$
$$\dot{\overline{x}} = \lceil A \rceil^{\mathrm{Mzl}} \overline{x} + \lfloor A \rfloor^{\mathrm{Mzl}} \underline{x} + B^{+} \overline{w} + B^{-} \underline{w}$$



Reachability using Embedding Systems

Example

Original System:

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} {\color{red}x_2^3} {\color{blue}- x_2} + {\color{red}w} \\ {\color{red}x_1} \end{bmatrix}$$

$$\mathcal{W} = [2.2\ ,\ 2.3] \quad \mathcal{X}_0 = \left[\begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix}, \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}\right]$$

red $=$ cooperative, blue $=$ competitive

Embedding System:

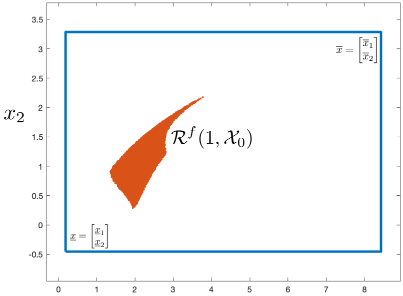
$$\frac{d}{dt}\begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \\ \overline{x}_1 \\ \overline{x}_2 \end{bmatrix} = \begin{bmatrix} \underline{x}_2^3 - \overline{x}_2 + \underline{w} \\ \underline{x}_1 \\ \overline{x}_2^3 - \underline{x}_2 + \overline{w} \\ \overline{x}_1 \end{bmatrix} \qquad \begin{bmatrix} \underline{w} \\ \overline{w} \end{bmatrix} = \begin{bmatrix} 2.2 \\ 2.3 \end{bmatrix}$$

$$\begin{bmatrix} \underline{x}_1(0) \\ \underline{x}_2(0) \end{bmatrix} = \begin{bmatrix} -0.5 \\ -0.5 \end{bmatrix} \qquad \begin{bmatrix} \overline{x}_1(0) \\ \overline{x}_2(0) \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$$

Decomposition function

$$\underline{d}(\underline{x}, \overline{x}, \underline{w}, \overline{w}) = \begin{bmatrix} \underline{x}_2^3 + \underline{w} \\ \underline{x}_1 \end{bmatrix} + \begin{bmatrix} -\overline{x}_2 \\ 0 \end{bmatrix}$$

$$\overline{d}(\underline{x}, \overline{x}, \underline{w}, \overline{w}) = \begin{bmatrix} \overline{x}_2^3 + \overline{w} \\ \overline{x}_1 \end{bmatrix} + \begin{bmatrix} -\underline{x}_2 \\ 0 \end{bmatrix}$$



Decomposition functions

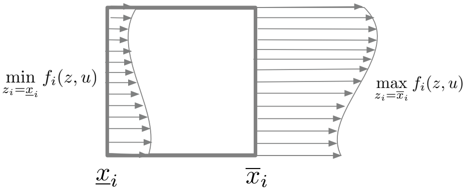
How to decompose a vector field?

Theorem

The best (tightest) decomposition function is given by

$$\underline{d}_i(\underline{x},\overline{x},\underline{w},\overline{w}) = \min_{\substack{z\in[\underline{x},\overline{x}],z_i=\underline{x}_i \\ u\in[\underline{w},\overline{w}]}} f_i(z,u),$$

$$\overline{d}_i(\underline{x},\overline{x},\underline{w},\overline{w}) = \max_{\substack{z\in[\underline{x},\overline{x}],z_i=\overline{x}_i \\ u\in[\underline{w},\overline{w}]}} f_i(z,u)$$



These optimization problems are, in general, not tractable

1. **Compositional approach:**
   - find the tight decomposition functions for a handful of elementary functions
   - write the vector field as composition of some elementary functions
2. **Jacobian-based approach:**
   - Taylor expansion for the vector field
   - formulas for linear systems

Tightness of Mixed Monotone Bounds

Diagonally weighted ℓ_∞-norm contractions

Theorem

Let $\frac{d}{dt}\begin{bmatrix}\underline{x}\\ \overline{x}\end{bmatrix} = \begin{bmatrix}\underline{d}(\underline{x},\overline{x},\underline{w},\overline{w})\\ \overline{d}(\underline{x},\overline{x},\underline{w},\overline{w})\end{bmatrix} := e(\underline{x},\overline{x},\underline{w},\overline{w})$ be the embedding function with the tight decomposition functions for $\dot{x} = f(x,w)$. For any $\eta \in \mathbb{R}^n_{\geq 0}$

$$\mu_{\infty,\eta}\left(\frac{\partial f}{\partial x}(x,w)\right) \leq c \qquad \text{for all } x, w$$

if and only if

$$\mu_{\infty,\eta\otimes I_2}\left(\frac{\partial e}{\partial[\frac{\underline{x}}{\overline{x}}]}(\underline{x},\overline{x},\underline{w},\overline{w})\right) \leq c, \qquad \text{for all } \underline{x},\overline{x},\underline{w},\overline{w}$$

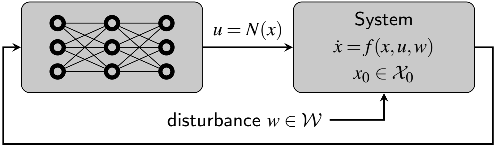
mixed monotone reachability is at least as accurate as contraction reachability wrt diagonally weighted ℓ_∞-norms

Given the open-loop nonlinear system with a neural network controller

$$\dot{x} = f^o(x, u, w),$$
$$u = N(x),$$

study reachability of the closed-loop system

$$\dot{x} = f^o(x, N(x), w) := f^c(x, w)$$



Challenge: finding a tight closed-loop decomposition function

Compositional Approach:

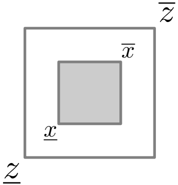
1. compute the open-loop decomposition function $\underline{d}^o, \overline{d}^o$ for $\dot{x} = f^o(x, u, w)$
2. construct the neural network bounding functions $\underline{N}$ and $\overline{N}$
   - bounds from NN verification algorithms (CROWN, LipSDP, IBP, etc)
   - construction is expensive but queries is cheap
3. combine (1) and (2) to obtain a decomposition function for the closed-loop system

Reachability of Neural Network Loops

Neural Network Bounds

Given a neural network controller $u = N(x)$ and an interval $[\underline{z}, \overline{z}]$,

$$\underline{N}_{[\underline{z},\overline{z}]}(\underline{x}, \overline{x}) \leq N(x) \leq \overline{N}_{[\underline{z},\overline{z}]}(\underline{x}, \overline{x}), \qquad \text{for all } [\underline{x}, \overline{x}] \subseteq [\underline{z}, \overline{z}]$$



- CROWN[2] provides affine bounds of the form

$$\underline{A}_{[\underline{z},\overline{z}]} x + \underline{b}_{[\underline{z},\overline{z}]} \leq N(x) \leq \overline{A}_{[\underline{z},\overline{z}]} x + \overline{b}_{[\underline{z},\overline{z}]}, \qquad \text{for all } x \in [\underline{z}, \overline{z}]. \tag{1}$$

Given the CROWN bounds (1), we have

$$\underline{N}_{[\underline{z},\overline{z}]}(\underline{x}, \overline{x}) = \underline{A}^{+}_{[\underline{z},\overline{z}]} \underline{x} + \overline{A}^{-}_{[\underline{z},\overline{z}]} \underline{x} + \underline{b}_{[\underline{z},\overline{z}]},$$

$$\overline{N}_{[\underline{z},\overline{z}]}(\underline{x}, \overline{x}) = \overline{A}^{+}_{[\underline{z},\overline{z}]} \overline{x} + \underline{A}^{-}_{[\underline{z},\overline{z}]} \underline{x} + \overline{b}_{[\underline{z},\overline{z}]}$$

[2] Zhang, Weng, Chen, Hsieh,Daniel, "Efficient neural network robustness certification with general activation functions." NeurIPS, 2018.

Theorem

Given open-loop decomposition functions $\underline{d}^o, \overline{d}^o$ for $\dot{x} = f^o(x, u, w)$ and the neural network bounds $\underline{N}_{[\underline{z},\overline{z}]}, \overline{N}_{[\underline{z},\overline{z}]}$ for the neural network controller $u = N(x)$. Then

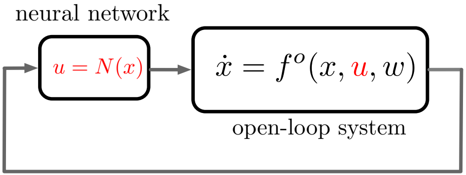
$$\underline{d}_i^c(\underline{x}, \overline{x}, \underline{w}, \overline{w}) = \underline{d}_i^o(\underline{x}, \overline{x}, \underline{\eta}^i, \overline{\eta}^i, \underline{w}, \overline{w})$$
$$\overline{d}_i^c(\underline{x}, \overline{x}, \underline{w}, \overline{w}) = \overline{d}_i^o(\underline{x}, \overline{x}, \underline{\nu}^i, \overline{\nu}^i, \underline{w}, \overline{w})$$

where

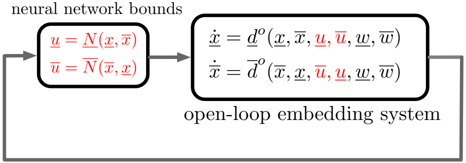
$$\underline{\eta}^i = \underline{N}_{[\underline{x},\overline{x}]}(\underline{x}, \overline{x}_{[i:\underline{x}]}) \quad \overline{\eta}^i = \overline{N}_{[\underline{x},\overline{x}]}(\underline{x}, \overline{x}_{[i:\underline{x}]}),$$
$$\underline{\nu}^i = \underline{N}_{[\underline{x},\overline{x}]}(\underline{x}_{[i:\overline{x}]}, \overline{x}) \quad \overline{\nu}^i = \overline{N}_{[\underline{x},\overline{x}]}(\underline{x}_{[i:\overline{x}]}, \overline{x}),$$

are decomposition functions for the closed-loop system where $v_{[i:w]}$ is the vector v with ith component replaced with ith component of w.

Original system:

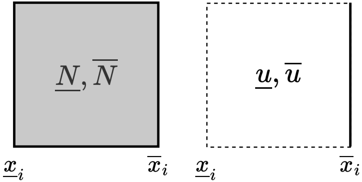


Embedding system:



How does the interconnection work?

- NN bounds $\underline{N}_{[\underline{x},\overline{x}]}, \overline{N}_{[\underline{x},\overline{x}]}$ are constructed on $[\underline{x}, \overline{x}]$
- NN bounds are evaluated on each edge, i.e., $\underline{u} = \underline{N}_{[\underline{x},\overline{x}]}(\underline{x}, \overline{x}_{[i:\underline{x}]})$ and $\overline{u} = \overline{N}_{[\underline{x},\overline{x}]}(\underline{x}, \overline{x}_{[i:\underline{x}]})$



Conclusions

- mixed monotone and contraction-based reachability
- connection between these two reachability approaches
- compositional approach for reachability analysis of NN controlled systems

Example: Bicycle Model

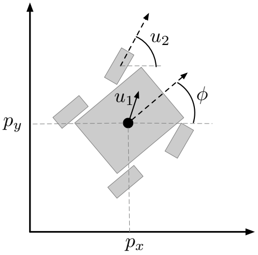
Design of the neural network

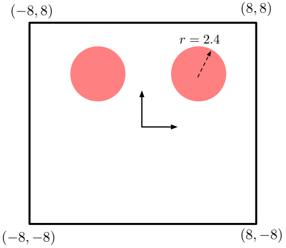
Dynamics of bicycle

$$\dot{p_x} = v\cos(\phi + \beta(u_2)) \qquad \dot{\phi} = \frac{v}{\ell_r}\sin(\beta(u_2))$$

$$\dot{p_y} = v\sin(\phi + \beta(u_2)) \qquad \dot{v} = u_1$$

$$\beta(u_2) = \arctan\left(\frac{l_r}{l_f + l_r}\tan(u_2)\right)$$





Goal: steer the bicycle to the origin avoiding the obstacles

- **offline controller:** MPC with hard constraint to avoid the obstacles
- run MPC for 65000 randomly chosen initial condition (20 sample per trajectory)
- train a feedforward neural network $4 \mapsto 100 \mapsto 100 \mapsto 2$ with this data

- start from $(8,8)$ toward $(0,0)$
- $\mathcal{X}_0 = [\underline{x}_0, \overline{x}_0]$ with

$$\underline{x}_0 = \begin{pmatrix} 7.9 & 7.9 & -\frac{2\pi}{3} - 0.01 & 1.99 \end{pmatrix}^\top$$
$$\overline{x}_0 = \begin{pmatrix} 8.1 & 8.1 & -\frac{2\pi}{3} + 0.01 & 2.01 \end{pmatrix}^\top$$

- CROWN for verification of neural network
- partition the states to improve accuracy
- mixed monotone approach **certify** that closed-loop system is avoiding the obstacle