Advanced Tools for Verification of Learning-Based Control

Bin Hu, Peter Seiler, Geir Dullerud

UIUC & UMich

AFOSR Meeting on Dynamical Systems and Control Theory 2024

Progress in Year 1

Our projects starts at 09/15/2023. In Year 1 (09/15/2023-now), we made significant progress in developing advanced tools for verification of learning-based control. Our main results include:

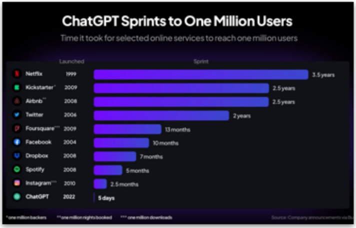
- S. Noori, B. Hu, G. Dullerud, P. Seiler. Stability and performance analysis of discrete-time ReLU recurrent neural networks, accepted to CDC, 2024. https://arxiv.org/abs/2405.05236
- S. Noori, B. Hu, G. Dullerud, P. Seiler. A complete set of quadratic constraints for repeated ReLU and generalizations, submitted to IEEE TAC, 2024. https://arxiv.org/abs/2407.06888
- A. Havens, P. Seiler, G. Dullerud, B. Hu. A quantitative local small gain theorem without gains, in preparation to IEEE TAC, 2024.

Some collaborative efforts on the machine learning side:

- Z. Wang, B. Hu, A. Havens, A. Araujo, Y. Zheng, Y. Chen, S. Jha. On the scalability and memory efficiency of semidefinite programs for Lipschitz constant estimation of neural networks, ICLR, 2024. https://openreview.net/forum?id=dwzLn78jq7
- A. Havens, A. Araujo, H. Zhang, B. Hu. Fine-grained local sensitivity analysis of standard dot-product self-attention. ICML, 2024. https://proceedings.mlr.press/v235/havens24a.html

Artificial Intelligence Revolution

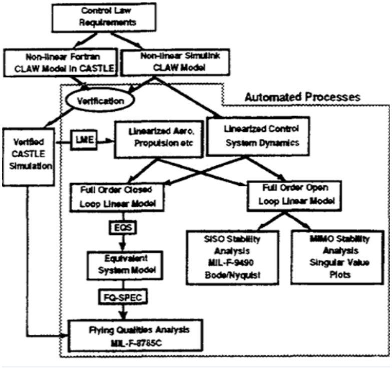
nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
At last — a computer program that can beat a champion Go player PAGE 484
ALL SYSTEMS GO

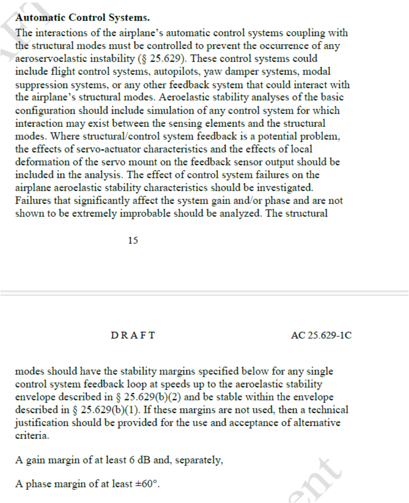


Deep learning models such as neural networks (NNs) and transformers have shown great promise for many tasks!

Safety-critical applications!

Verification is Crucial!


Control Law Requirements
Non-linear Fortran CLAW Model in CASTLE
Non-linear Simulink CLAW Model
Automated Processes
Verification
Verified CASTLE Simulation
LME
Linearized Aero, Propulsion etc
Linearized Control System Dynamics
Full Order Closed Loop Linear Model
Full Order Open Loop Linear Model
EQS
Equivalent System Model
FQ-SPEC
SISO Stability Analysis MIL-F-9490 Bode/Nyquist
MIMO Stability Analysis Singular Value Plots
Flying Qualities Analysis MIL-F-8785C



Automatic Control Systems.

The interactions of the airplane's automatic control systems coupling with the structural modes must be controlled to prevent the occurrence of any aeroservoelastic instability (§ 25.629). These control systems could include flight control systems, autopilots, yaw damper systems, modal suppression systems, or any other feedback system that could interact with the airplane's structural modes. Aeroelastic stability analyses of the basic configuration should include simulation of any control system for which interaction may exist between the sensing elements and the structural modes. Where structural/control system feedback is a potential problem, the effects of servo-actuator characteristics and the effects of local deformation of the servo mount on the feedback sensor output should be included in the analysis. The effect of control system failures on the airplane aeroelastic stability characteristics should be investigated. Failures that significantly affect the system gain and/or phase and are not shown to be extremely improbable should be analyzed. The structural

15

DRAFT AC 25.629-1C

modes should have the stability margins specified below for any single control system feedback loop at speeds up to the aeroelastic stability envelope described in § 25.629(b)(2) and be stable within the envelope described in § 25.629(b)(1). If these margins are not used, then a technical justification should be provided for the use and acceptance of alternative criteria.

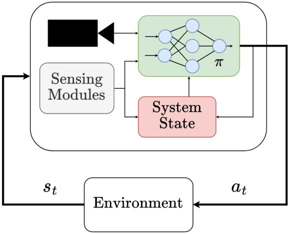
A gain margin of at least 6 dB and, separately,

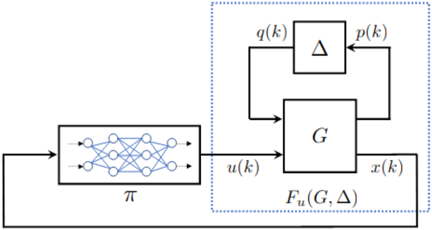
A phase margin of at least ±60°.

Our research aims at bridging the gap between modern deep learning models and the verification requirements on safety-critical systems!

Robustness Verification of Learning-Based Control

We are interested in robustness verification of learning-based control with deep learning components in the loop:





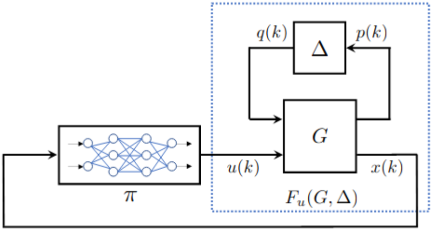
- Δ can be dynamical uncertainty (with unknown order) or time-varying delay
- The controller π is a deep learning model
- Complex perception errors due to high-dimensional sensory data
- **Issues**: Conservatism, generality, and scalability of verification

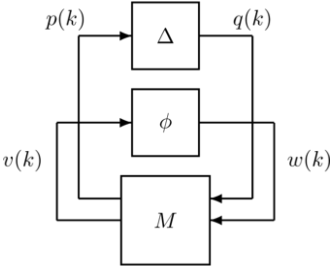
Outline

- Mitigating Conservatism via Complete Quadratic Constraints
- Improving Generality: Robustness Analysis for Transformers
- Plan for Year 2

Quadratic Constraints for NN Controllers

The quadratic constraint (QC) approach can address neural networks and dynamical uncertainty simultaneously for verifying robust region of attractions:



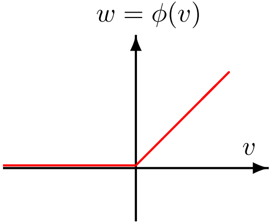


- Key idea (Fazlyab, Morari, Pappas 2019; Yin, Seiler, Arcak 2021): Rewrite the original system as a feedback loop of LTI M and troublesome elements (Δ, ϕ) and then abstract the troublesome elements using QCs
- Nonlinearity ϕ is the activation function used in neural networks
- Dynamic integral quadratic constraints for dynamical uncertainty Δ
- Firmly connected to dissipativity theory

Repeated Rectified Linear Unit (ReLU)

The scalar Rectified Linear Unit (ReLU) is a function $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ defined by:

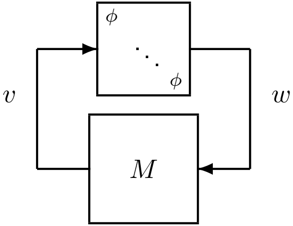
$$\phi(v) = \begin{cases} 0 & \text{if } v < 0 \\ v & \text{if } v \geq 0 \end{cases} .$$



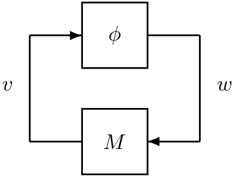
i) Positivity: $\phi(v) \geq 0$. *ii) Positive Complement:* $\phi(v) \geq v$. *iii) Complementarity:* $\phi(v)\,(v - \phi(v)) = 0$. *iv) Positive Homogeneity:* $\phi(\beta v) = \beta \phi(v)\ \forall v \in \mathbb{R}, \beta \geq 0$. *v) Slope-restricted on* $[0,1]$*:* $0 \leq \frac{\phi(v)-\phi(\tilde{v})}{v-\tilde{v}} \leq 1$

The repeated ReLU is the function mapping from $\mathbb{R}^{n_v}$ to $\mathbb{R}^{n_v}_{\geq 0}$ defined by

$$w := \phi(v) = \begin{bmatrix} \phi(v_1) \\ \phi(v_2) \\ \vdots \\ \phi(v_{n_v}) \end{bmatrix}$$



Review: Basic Ideas of QCs for NN Controllers



- **Goal:** Analyze the following set of coupled sequences $\{\xi(k), w(k), v(k)\}$

$$\{(\xi, w, v) : \xi(k+1) = A\xi(k) + Bw(k),\ v(k) = C\xi(k)\} \cap \{(\xi, w, v) : w(k) = \phi(v(k))\}$$

- **Key idea: Quadratic constraints!** Replace the graph of ϕ with a relation on (v, w) captured by a quadratic constraint:

$$\{(v, w) : w(k) = \phi(v(k))\} \subset \left\{(v, w) : \begin{bmatrix} v(k) \\ w(k) \end{bmatrix}^{\mathsf{T}} \Lambda \begin{bmatrix} v(k) \\ w(k) \end{bmatrix} \geq 0\right\},$$

  where Λ is constructed from the property of ϕ.

- We only need to analyze the stability of the following set:

$$\left\{(\xi, w, v) : \xi(k+1) = A\xi(k) + Bw(k),\ v(k) = C\xi(k), \begin{bmatrix} v(k) \\ w(k) \end{bmatrix}^{\mathsf{T}} \Lambda \begin{bmatrix} v(k) \\ w(k) \end{bmatrix} \geq 0\right\}.$$

Review: Basic Ideas of QCs for NN Controllers

Now we are analyzing the sequence from the following set:

$$\left\{(\xi, w, v) : \xi(k+1) = A\xi(k) + Bw(k),\, v(k) = C\xi(k),\, \begin{bmatrix} v(k) \\ w(k) \end{bmatrix}^\mathsf{T} \Lambda \begin{bmatrix} v(k) \\ w(k) \end{bmatrix} \geq 0\right\}.$$

Theorem

If there exists a positive definite matrix P and $0 < \rho < 1$ s.t.

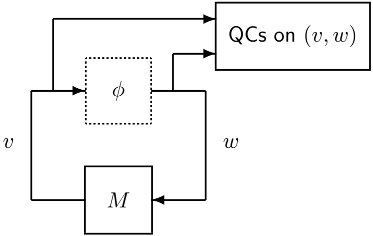
$$\begin{bmatrix} A^\mathsf{T} PA - \rho^2 P & A^\mathsf{T} PB \\ B^\mathsf{T} PA & B^\mathsf{T} PB \end{bmatrix} \preceq - \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix}^\mathsf{T} \Lambda \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix}$$

then $\xi(k+1)^\mathsf{T} P\xi(k+1) \leq \rho^2 \xi(k)^\mathsf{T} P\xi(k)$ *and* $\lim_{k\to\infty} \xi(k) = 0$.

$$\underbrace{\begin{bmatrix} \xi(k) \\ w(k) \end{bmatrix}^\mathsf{T} \begin{bmatrix} A^\mathsf{T} PA - \rho^2 P & A^\mathsf{T} PB \\ B^\mathsf{T} PA & B^\mathsf{T} PB \end{bmatrix} \begin{bmatrix} \xi(k) \\ w(k) \end{bmatrix}}_{\xi(k+1)^\mathsf{T} P\xi(k+1) - \rho^2 \xi(k)^\mathsf{T} P\xi(k)} \leq \underbrace{- \begin{bmatrix} \xi(k) \\ w(k) \end{bmatrix}^\mathsf{T} \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix}^\mathsf{T} \Lambda \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \xi(k) \\ w(k) \end{bmatrix}}_{- \begin{bmatrix} v(k) \\ w(k) \end{bmatrix}^\mathsf{T} \Lambda \begin{bmatrix} v(k) \\ w(k) \end{bmatrix} \leq 0}$$

This condition is a semidefinite program (SDP) problem!

QCs for Feedback Systems with ReLU Networks



- To reduce conservatism, replace troublesome ϕ with multiple QCs:

$$\{(v,w) : w(k) = \phi(v(k))\} \subset \bigcap_{\Lambda \in \mathcal{M}} \left\{ (v,w) : \begin{bmatrix} v(k) \\ w(k) \end{bmatrix}^\mathsf{T} \Lambda \begin{bmatrix} v(k) \\ w(k) \end{bmatrix} \geq 0 \right\},$$

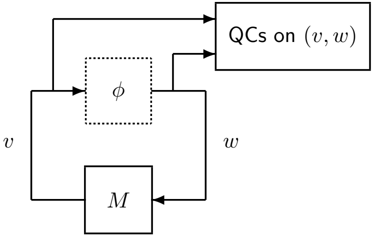
- [Willems, Brockett, '68; Willems '71]: If Q_0 is doubly hyperdominant, then

$$\begin{bmatrix} v \\ w \end{bmatrix}^\mathsf{T} \begin{bmatrix} 0 & Q_0^\mathsf{T} \\ Q_0 & -(Q_0 + Q_0^\mathsf{T}) \end{bmatrix} \begin{bmatrix} v \\ w \end{bmatrix} \geq 0$$

- [Ebihara, et al '21, Richardson, et al '23]: Many papers on QCs for ReLU!
- **Advantages:** General stability/robustness/input-output analysis

Ref: Noori, Hu, Dullerud, Seiler (CDC 2024)

Conservatism in QCs for ReLU Networks



- To reduce conservatism, replace troublesome ϕ with multiple QCs:

$$\{(v,w) : w(k) = \phi(v(k))\} \subset \bigcap_{\Lambda \in \mathcal{M}} \left\{ (v,w) : \begin{bmatrix} v(k) \\ w(k) \end{bmatrix}^\mathsf{T} \Lambda \begin{bmatrix} v(k) \\ w(k) \end{bmatrix} \geq 0 \right\},$$

- Issue: If the set on the right side is much **larger** than the set on the left side, then the analysis can become conservative!
- **Our main result:** The complete set $\mathcal{M}$ of QCs on repeated ReLU can be derived. The use of the complete set $\mathcal{M}$ does not introduce conservatism in a formal sense.

Complete QCs for Repeated ReLU

A fundamental question: Have we found the complete set of quadratic constraints on repeated ReLU? In other words, have we found all the matrices $\Lambda \in \mathbb{R}^{2n_v \times 2n_v}$ which can guarantee the following inequality with ϕ being repeated ReLU?

$$\begin{bmatrix} w \\ v \end{bmatrix}^\mathsf{T} \Lambda \begin{bmatrix} w \\ v \end{bmatrix} \geq 0, \forall w = \phi(v)$$

The answer is NO!

Theorem (NHDS2024)

$\Lambda \in \mathbb{R}^{2n_v \times 2n_v}$ *gives a valid QC for repeated ReLU if and only if* $\begin{bmatrix} D \\ \frac{1}{2}(I + D) \end{bmatrix}^\mathsf{T} \Lambda \begin{bmatrix} D \\ \frac{1}{2}(I + D) \end{bmatrix}$ *is copositive* $\forall D \in \mathrm{diag}(\{1, -1\}^{n_v})$.

Ref: Noori, Hu, Dullerud, Seiler (Submitted to TAC, 2024) https://arxiv.org/abs/2407.06888

Complete QCs for Repeated ReLU

Theorem (NHDS2024)

$\Lambda \in \mathbb{R}^{2n_v \times 2n_v}$ *gives a valid QC for repeated ReLU if and only if* $\begin{bmatrix} D \\ \frac{1}{2}(I+D) \end{bmatrix}^{\mathsf{T}} \Lambda \begin{bmatrix} D \\ \frac{1}{2}(I+D) \end{bmatrix}$ *is copositive* $\forall D \in \operatorname{diag}(\{1,-1\}^{n_v})$.

- Q is **copositive** if $x^{\mathsf{T}} Q x \geq 0$ for all x with only non-negative entries
- All existing QCs on repeated ReLU can be re-derived via the above theorem.
- Proof idea: The graph of the repeated ReLU satisfies:

$$\left\{ \begin{bmatrix} v \\ w \end{bmatrix} \in \mathbb{R}^{2n_v} \middle| v \in \mathbb{R}^{n_v} \text{ and } w = \phi(v) \right\}$$
$$= \left\{ \begin{bmatrix} D \\ \frac{1}{2}(I+D) \end{bmatrix} \tilde{v} \middle| \tilde{v} \in \mathbb{R}_+^{n_v}, D \in \operatorname{diag}(\{1,-1\}^{n_v}) \right\}$$

In other words, we write $v = D\tilde{v}$ where $\tilde{v} = |v|$ and $D = \operatorname{diag}(\operatorname{sign}(v))$. Then entry k of $\frac{1}{2}(I+D)\tilde{v}$ is 0 if $D_{kk} = -1$ or equal to the entry k of v if $D_{kk} = 1$.

Complete QCs for Repeated ReLU

Denote $\mathcal{M}$ to be the set of all Λ such that

$$\begin{bmatrix} D \\ \frac{1}{2}(I+D) \end{bmatrix}^\mathsf{T} \Lambda \begin{bmatrix} D \\ \frac{1}{2}(I+D) \end{bmatrix} \text{ is copositive } \forall D \in \operatorname{diag}(\{1,-1\}^{n_v}).$$

Theorem (NHDS2024)

A function satisfies all the QCs defined by the previous complete set $\mathcal{M}$ if and only if the function is either repeated ReLU or flipped ReLU.

- Our complete set of QCs is as tight as possible up to the sign invariance inherent in quadratic forms.
- The complete set $\mathcal{M}$ does not introduce conservatism in the following sense:

$$\begin{aligned} &\{(v,w) : w(k) = \phi(v(k)) \text{ or } w(k) = -\phi(-v(k))\} \\ &= \bigcap_{\Lambda \in \mathcal{M}} \left\{ (v,w) : \begin{bmatrix} v(k) \\ w(k) \end{bmatrix}^\mathsf{T} \Lambda \begin{bmatrix} v(k) \\ w(k) \end{bmatrix} \geq 0 \right\}. \end{aligned}$$

Complete Incremental QCs for Lipschitz Constant Analysis of NNs

Incremental QCs are crucial for Lipschitz analysis of NNs [Fazlyab et.al 2019]:

$$\begin{bmatrix} v-\tilde{v} \\ \phi(v)-\phi(\tilde{v}) \end{bmatrix}^\mathsf{T} \Lambda \begin{bmatrix} v-\tilde{v} \\ \phi(v)-\phi(\tilde{v}) \end{bmatrix} \geq 0, \forall v, \tilde{v} \in \mathbb{R}^{n_v}$$

Theorem (NHDS2024)

$\Lambda \in \mathbb{R}^{2n_v \times 2n_v}$ *gives an incremental QC for repeated ReLU if & only if*

$$\begin{bmatrix} D_1 & -D_2 \\ \frac{1}{2}(I+D_1) & -\frac{1}{2}(I+D_2) \end{bmatrix}^\mathsf{T} \Lambda \begin{bmatrix} D_1 & -D_2 \\ \frac{1}{2}(I+D_1) & -\frac{1}{2}(I+D_2) \end{bmatrix}$$

is copositive $\forall D_1, D_2 \in \mathrm{diag}(\{1,-1\}^{n_v})$.

- All existing incremental QCs on repeated ReLU can be re-derived via the above theorem.
- The above theorem provides a unified approach for incremental QCs.
- The above theorem can improve Lipschitz analysis of ReLU networks.

Generalizations to HouseHolder and MaxMin

We can also obtain complete QCs for other piecewise linear activation functions.

- Leaky ReLU: Define $g_{\alpha\beta} : \mathbb{R} \to \mathbb{R}$ for $\alpha \neq \beta$ as follows:

  $$g_{\alpha\beta}(v) = \begin{cases} \alpha v & \text{if } v < 0 \\ \beta v & \text{if } v \geq 0 \end{cases}.$$

  $g_{\alpha\beta}$ is the scalar ReLU when $\alpha = 0$ and $\beta = 1$. It corresponds to leaky ReLU when $0 < \alpha < 1$ and $\beta = 1$. We can show the complete QC set for leaky ReLU is given by $\begin{bmatrix} D \\ \alpha D + \frac{\beta-\alpha}{2}(I+D) \end{bmatrix}^\mathsf{T} \Lambda \begin{bmatrix} D \\ \alpha D + \frac{\beta-\alpha}{2}(I+D) \end{bmatrix}$ being copositive $\forall D \in \operatorname{diag}(\{1,-1\}^{n_v})$.

- HouseHolder/MaxMin: Given h with $\|h\|_2 = 1$, Householder is defined by:

  $$G_h(v) = \begin{cases} v & \text{if } h^\mathsf{T} v \geq 0 \\ (I - 2hh^\mathsf{T})v & \text{if } h^\mathsf{T} v < 0 \end{cases}.$$

  If $h = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \end{bmatrix}^\mathsf{T}$ then we have MaxMin activations. Both are widely used to improve certified robustness of neural network image classifiers.

  Ref: C. Anil, J. Lucas, and R. Grosse, Sorting out Lipschitz function approximation, ICML 2019.

  We have obtained complete sets for both!

Numerical Results

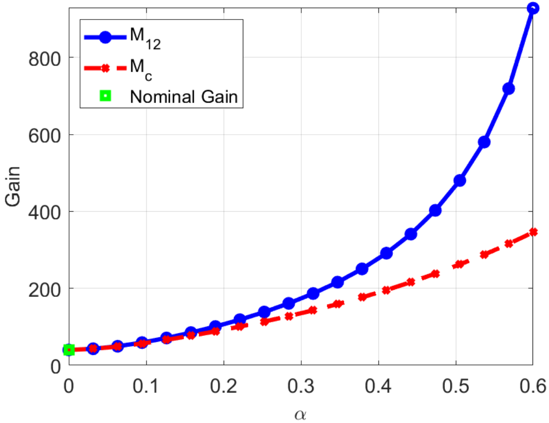


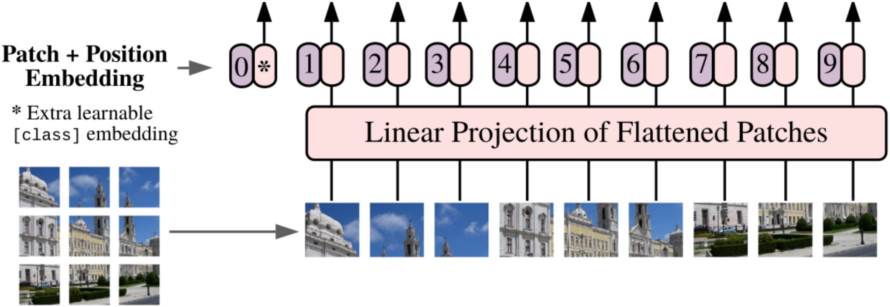
Figure: Bound on induced ℓ_2 gain for system $F_U(G,\Phi)$ vs. α using QCs defined by $\mathcal{M}_c$ and $\mathcal{M}_{12}$. We expect the interconnection to eventually become unstable as α increases which is consistent with both curves. The complete set $\mathcal{M}_c$ provides a less conservative (smaller) bound on the gain.

Outline

- Mitigating Conservatism via Complete Quadratic Constraints
- Improving Generality: Robustness Analysis for Transformers
- Plan for Year 2

Self-Attention and Transformers

The self-attention mechanism has become a major building block in many modern deep learning-based system, in particular Transformers



Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Figure: Alexey Dosovitskiy et al. 2021

Dot-Product Self-Attention

The main building-block of transformers are self-attention units. The dot-product multi-head self-attention maps $\mathbb{R}^{n\times d}$ to $\mathbb{R}^{n\times d}$. With h heads, the l-th head maps $\mathbb{R}^{n\times d}$ to $\mathbb{R}^{n\times d/h}$ as:

$$X = \begin{bmatrix} - & x_1^\intercal & - \\ & \vdots & \\ - & x_n^\intercal & - \end{bmatrix} \in \mathbb{R}^{n\times d} \qquad Y_l = \underbrace{\text{softmax}\left(\frac{XW_l^Q(XW_l^K)^\intercal}{\sqrt{d/h}}\right)}_{=P_l(X)} XW_l^V$$

where $W_l^Q, W_l^K, W_l^V \in \mathbb{R}^{d\times d/h}$ denote the weight matrices for the l-th head

- Issue 1: Dot-product self-attention is not globally Lipschitz [Kim et.al. ICML2021] and does not have a global gain bound.
- Issue 2: One cannot rewrite dot-product self-attention as a feedback interconnection of a LTI component and a nonlinearity

Key question: How to certify closed-loop robustness of such transformer components?

Robustness Analysis of Dot-Product Self-Attention

High-level idea: We decouple the analysis into two steps.

- Local sensitivity analysis of self-attention: Given an input X and some $\varepsilon > 0$, we want to prove a bound in the following form:

  $$\|F(X') - F(X)\|_F \leq \delta(X, \varepsilon) \quad \text{for } X' \text{ satisfying} \quad \|X' - X\|_F \leq \varepsilon$$

  where F is the self-attention mapping.

  Ref: A. Havens, A. Araujo, H. Zhang, B. Hu. Fine-grained local sensitivity analysis of standard dot-product self-attention. ICML, 2024.

- A gainless version of small gain theorem (which incorporates a local sensitivity bound into the closed-loop robustness property)

  Ref: A. Havens, P. Seiler, G. Dullerud, B. Hu. A quantitative local small gain theorem without gains, in preparation to IEEE TAC, 2024.

Local Sensitivity of Transformers

$$\begin{aligned}
&\|F(X') - F(X)\|_F \\
=&\Big\|H(X'-X) + \sum_{l=1}^{h} P_l(X)(X'-X)W_l^V W_l^O + \sum_{l=1}^{h}(P_l(X') - P_l(X))X'W_l^V W_l^O\Big\|_F \\
\le& \underbrace{\Big\|H(X'-X) + \sum_{l=1}^{h} P_l(X)(X'-X)W_l^V W_l^O\Big\|_F}_{\Delta_1 \le \delta_1(X,\varepsilon)} + \underbrace{\Big\|\sum_{l=1}^{h}(P_l(X') - P_l(X))X'W_l^V W_l^O\Big\|_F}_{\Delta_2 \le \delta_2(X,\varepsilon)}
\end{aligned}$$

$$\boxed{\max_{X':\|X'-X\|_F \le \varepsilon} \|F(X') - F(X)\|_F \le \delta_1(X,\varepsilon) + \delta_2(X,\varepsilon)}$$

We bound $\delta_1(X,\varepsilon)$ as $\delta_1(X,\varepsilon) = \left\|H \otimes I_n + \sum_{l=1}^{h}(P_l(X) \otimes (W_l^V W_l^O)^\mathsf{T})\right\|\varepsilon$

The calculation of $\delta_2(X,\varepsilon)$ is complicated:

$$\begin{aligned}
\delta_2(X,\varepsilon) = \frac{\varepsilon}{\sqrt{d/h}} \sum_{l=1}^{h} &\left(\|XW_l^V W_l^O\| + \|W_l^V W_l^O\|\varepsilon\right) \cdot \\
&\Big(\|W_l^Q (W_l^K)^\mathsf{T} X^\mathsf{T}\| + \|XW_l^Q (W_l^K)^\mathsf{T}\| + \varepsilon\|W_l^Q (W_l^K)^\mathsf{T}\|\Big)
\end{aligned}$$

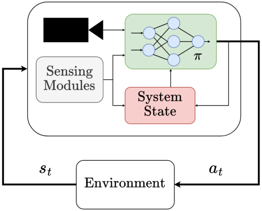
The final bound is much tighter than doing a local Lipschitz analysis.

Outline

- Mitigating Conservatism via Complete Quadratic Constraints
- Improving Generality: Robustness Analysis for Transformers
- Plan for Year 2

Plan for Year 2

1. **Scalability:** We are studying how to scale up the SDP for complete QCs for practical deep learning models.
   - One plausible approach is to derive equivalent unconstrained formulation of the original SDP condition derived from complete QCs.
   - Initial results for scaling up Lipschitz analysis to ImageNet
     Ref: Z. Wang, B. Hu, A. Havens, A. Araujo, Y. Zheng, Y. Chen, S. Jha. On the scalability and memory efficiency of semidefinite programs for Lipschitz constant estimation of neural networks, ICLR, 2024.
   - Next: Scale up the computation for verifying closed-loop robustness!
2. **Verifying perception-based control:** We will integrate perception errors into the verification framework.



Question: How can we characterize perception errors in a form that is friendly for verification?