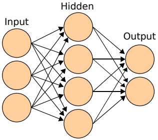
Polynomial Optimization for Bounding Lipschitz Constants of Deep Networks

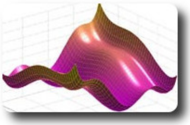
Victor Magron, MAC Team, CNRS–LAAS

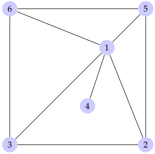
Jointly certified with T. Chen, J.-B. Lasserre and E. Pauwels

IPAM, UCLA

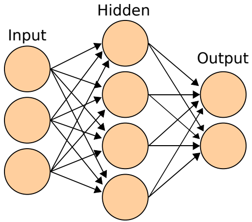
28 February 2020







Lipschitz constant of neural networks



- Applications: WGAN, certification
- Existing works: [Lattore et al.'18] based on linear programming (LP)
- Network setting: K-classifier, **ReLU network**, $1 + m$ layers (1 input layer + m hidden layer), $\mathbf{A}_i$ weights, $\mathbf{b}_i$ biases
- Score of label $k \leqslant K$ = ${\mathbf{c}_k}^T \mathbf{x}_m$ with last activation vector $\mathbf{c}_k$

Lipschitz constant of neural networks

$$\mathbf{x}_0 \in \mathbb{R}^{p} \longrightarrow \mathbf{z}_0 \in \mathbb{R}^{p} \longrightarrow \mathbf{z}_1 \in \mathbb{R}^{p_1} \longrightarrow \cdots \longrightarrow \mathbf{z}_m \in \mathbb{R}^{p_m}$$

$$\mathbf{z}_i = \mathbf{A}_i \mathbf{x}_{i-1} + \mathbf{b}_i \qquad \mathbf{x}_{i-1} = \mathrm{ReLU}(\mathbf{z}_{i-1})$$

Lipschitz constant of neural networks

$$\mathbf{x}_0 \in \mathbb{R}^p \longrightarrow \mathbf{z}_0 \in \mathbb{R}^p \longrightarrow \mathbf{z}_1 \in \mathbb{R}^{p_1} \longrightarrow \cdots \longrightarrow \mathbf{z}_m \in \mathbb{R}^{p_m}$$

$$\mathbf{z}_i = \mathbf{A}_i \mathbf{x}_{i-1} + \mathbf{b}_i \qquad \mathbf{x}_{i-1} = \mathrm{ReLU}(\mathbf{z}_{i-1})$$

LIPSCHITZ CONSTANT:

$$L_f^{||\cdot||} = \inf\{L : \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}, |f(\mathbf{x}) - f(\mathbf{y})| \leq L||\mathbf{x} - \mathbf{y}||\}$$

Lipschitz constant of neural networks

$$\mathbf{x}_0 \in \mathbb{R}^{p} \longrightarrow \mathbf{z}_0 \in \mathbb{R}^{p} \longrightarrow \mathbf{z}_1 \in \mathbb{R}^{p_1} \longrightarrow \cdots \longrightarrow \mathbf{z}_m \in \mathbb{R}^{p_m}$$

$$\mathbf{z}_i = \mathbf{A}_i \mathbf{x}_{i-1} + \mathbf{b}_i \qquad \mathbf{x}_{i-1} = \mathrm{ReLU}(\mathbf{z}_{i-1})$$

LIPSCHITZ CONSTANT:

$$\begin{aligned} L_f^{||\cdot||} &= \inf\{L : \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}, |f(\mathbf{x}) - f(\mathbf{y})| \leq L||\mathbf{x} - \mathbf{y}||\} \\ &= \sup\{||\nabla f(\mathbf{x})||_* : \mathbf{x} \in \mathcal{X}\} \\ &= \sup\{\mathbf{t}^T \nabla f(\mathbf{x}) : \mathbf{x} \in \mathcal{X}, ||\mathbf{t}|| \leq 1\} \end{aligned}$$

Lipschitz constant of neural networks

$$\mathbf{x}_0 \in \mathbb{R}^p \longrightarrow \mathbf{z}_0 \in \mathbb{R}^p \longrightarrow \mathbf{z}_1 \in \mathbb{R}^{p_1} \longrightarrow \cdots \longrightarrow \mathbf{z}_m \in \mathbb{R}^{p_m}$$

$$\mathbf{z}_i = \mathbf{A}_i \mathbf{x}_{i-1} + \mathbf{b}_i \qquad \mathbf{x}_{i-1} = \mathrm{ReLU}(\mathbf{z}_{i-1})$$

LIPSCHITZ CONSTANT:

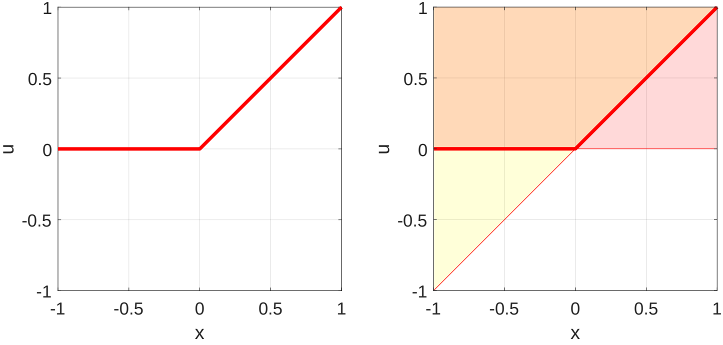
$$\begin{aligned} L_f^{||\cdot||} &= \inf\{L : \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}, |f(\mathbf{x}) - f(\mathbf{y})| \leq L||\mathbf{x} - \mathbf{y}||\} \\ &= \sup\{||\nabla f(\mathbf{x})||_* : \mathbf{x} \in \mathcal{X}\} \\ &= \sup\{\mathbf{t}^T \nabla f(\mathbf{x}) : \mathbf{x} \in \mathcal{X}, ||\mathbf{t}|| \leq 1\} \end{aligned}$$

GRADIENT for a fixed label k:

$$\nabla f(\mathbf{x}_0) = \left(\prod_{i=1}^{m} \mathbf{A}_i{}^T \mathrm{diag}\,(\mathrm{ReLU}'(\mathbf{z}_i)) \right) \mathbf{c}_k$$

A polynomial optimization formulation

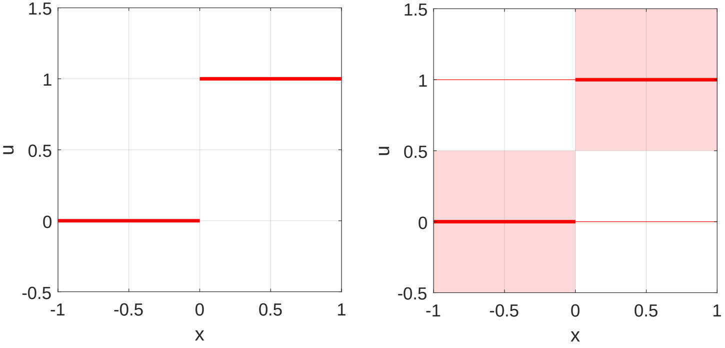
ReLU (left) & its semialgebraicity (right)



$$u = \max\{x, 0\} \qquad u(u - x) = 0, u \geq x, u \geq 0$$

A polynomial optimization formulation

ReLU' (left) & its semialgebraicity (right)



$u = \mathbf{1}_{\{x \geq 0\}}$ $\qquad$ $u(u-1) = 0, (u - \frac{1}{2})x \geq 0$

A polynomial optimization formulation

Local Lipschitz constant: $\mathbf{x}_0 \in$ ball of center $\bar{\mathbf{x}}_0$ and radius ε

A polynomial optimization formulation

Local Lipschitz constant: $\mathbf{x}_0 \in$ ball of center $\bar{\mathbf{x}}_0$ and radius ε

One single hidden layer ($m = 1$):

$$\sup_{\mathbf{x},\mathbf{u},\mathbf{z},\mathbf{t}} \mathbf{t}^T \mathbf{A}^T \text{diag}\,(\mathbf{u})\mathbf{c}$$

$$\text{s.t.} \begin{cases} (\mathbf{z} - \mathbf{A}\mathbf{x} - \mathbf{b})^2 = 0 \\ \mathbf{t}^2 \leq 1, (\mathbf{x} - \bar{\mathbf{x}}_0 + \varepsilon)(\mathbf{x} - \bar{\mathbf{x}}_0 - \varepsilon) \leq 0 \\ \mathbf{u}(\mathbf{u} - 1) = 0, (\mathbf{u} - 1/2)\mathbf{z} \geq 0 \end{cases}$$

A polynomial optimization formulation

Local Lipschitz constant: $\mathbf{x}_0 \in$ ball of center $\bar{\mathbf{x}}_0$ and radius ε

One single hidden layer ($m = 1$):

$$\sup_{\mathbf{x},\mathbf{u},\mathbf{z},\mathbf{t}} \mathbf{t}^T \mathbf{A}^T \text{diag}\,(\mathbf{u})\mathbf{c}$$

$$\text{s.t.} \begin{cases} (\mathbf{z} - \mathbf{A}\mathbf{x} - \mathbf{b})^2 = 0 \\ \mathbf{t}^2 \leq 1, (\mathbf{x} - \bar{\mathbf{x}}_0 + \varepsilon)(\mathbf{x} - \bar{\mathbf{x}}_0 - \varepsilon) \leq 0 \\ \mathbf{u}(\mathbf{u} - 1) = 0, (\mathbf{u} - 1/2)\mathbf{z} \geq 0 \end{cases}$$

“CHEAP” and “TIGHT” upper bound?

The moment-sums of squares hierarchy

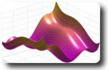
NP-hard NON CONVEX Problem $f^\star = \sup f(\mathbf{x})$

Theory

(Primal)

$\sup \int f \, d\mu$

with μ proba $\Rightarrow$



INFINITE LP

(Dual)

$\inf \lambda$

$\Leftarrow$ with $\lambda - f \geqslant 0$

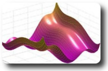
The moment-sums of squares hierarchy

NP-hard NON CONVEX Problem $f^\star = \sup f(\mathbf{x})$

Practice

(Primal **Relaxation**)

moments $\int \mathbf{x}^\alpha \, d\mu$



(Dual **Strengthening**)

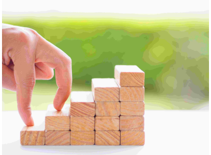
$\lambda - f =$ **sum of squares**

finite number $\Rightarrow$ **SDP** $\Leftarrow$ **fixed** degree

LASSERRE'S HIERARCHY of **CONVEX PROBLEMS** $\uparrow f^*$ [Lasserre/Parrilo 01]

degree d & n vars $\Longrightarrow \binom{n+2d}{n}$ **SDP** VARIABLES

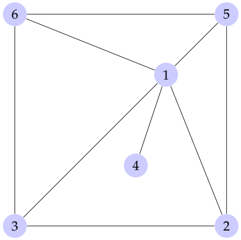
Numeric solvers $\Longrightarrow$ **Approx** Certificate



The sparse hierarchy [Waki, Lasserre 06]

- Correlative sparsity pattern

$$f = x_2x_5 + x_3x_6 - x_2x_3 - x_5x_6 + x_1(-x_1 + x_2 + x_3 - x_4 + x_5 + x_6)$$

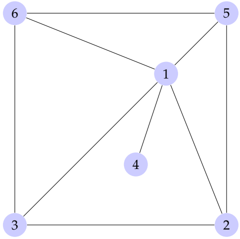


Chordal graph

The sparse hierarchy [Waki, Lasserre 06]

- Correlative sparsity pattern

$$f = x_2x_5 + x_3x_6 - x_2x_3 - x_5x_6 + x_1(-x_1 + x_2 + x_3 - x_4 + x_5 + x_6)$$



Chordal graph

$C_1 = \{1,4\}$

$C_2 = \{1,2,3,5\}$

$C_3 = \{1,3,5,6\}$

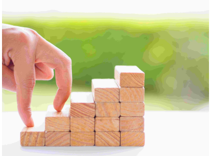
Dense SDP: 210 vars

Sparse SDP: 115 vars

1. Subsets C_1, C_2, C_3
2. Average size $\kappa \rightsquigarrow \binom{\kappa+2d}{\kappa}$ vars

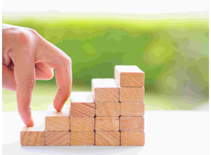
Our "heuristic relaxation" method: HR-2

💡 Go between 1ST & 2ND stair in SPARSE hierarchy



Our "heuristic relaxation" method: HR-2

Go between 1ST & 2ND stair in SPARSE hierarchy

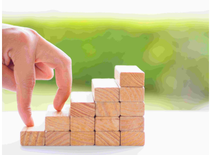


$$\sup_{\mathbf{x},\mathbf{u},\mathbf{z},\mathbf{t}} \mathbf{t}^T \mathbf{A}^T \text{diag}(\mathbf{u})\mathbf{c}$$

$$\text{s.t.} \begin{cases} (\mathbf{z} - \mathbf{A}\mathbf{x} - \mathbf{b})^2 = 0 \\ \mathbf{t}^2 \leq 1, (\mathbf{x} - \bar{\mathbf{x}}_0 + \varepsilon)(\mathbf{x} - \bar{\mathbf{x}}_0 - \varepsilon) \leq 0 \\ \mathbf{u}(\mathbf{u} - 1) = 0, (\mathbf{u} - 1/2)\mathbf{z} \geq 0 \end{cases}$$

Our "heuristic relaxation" method: HR-2

- Go between 1ST & 2ND stair in SPARSE hierarchy



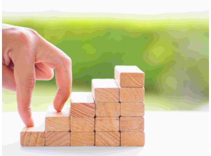
$$\sup_{\mathbf{x},\mathbf{u},\mathbf{z},\mathbf{t}} \mathbf{t}^T \mathbf{A}^T \text{diag}\,(\mathbf{u})\mathbf{c}$$

$$\text{s.t.} \begin{cases} (\mathbf{z} - \mathbf{A}\mathbf{x} - \mathbf{b})^2 = 0 \\ \mathbf{t}^2 \leq 1, (\mathbf{x} - \bar{\mathbf{x}}_0 + \varepsilon)(\mathbf{x} - \bar{\mathbf{x}}_0 - \varepsilon) \leq 0 \\ \mathbf{u}(\mathbf{u} - 1) = 0, (\mathbf{u} - 1/2)\mathbf{z} \geq 0 \end{cases}$$

- Pick SDP variables for products in $\{x, t\}$, $\{u, z\}$ up to deg 4

Our "heuristic relaxation" method: HR-2

- Go between 1ST & 2ND stair in SPARSE hierarchy



$$\sup_{\mathbf{x},\mathbf{u},\mathbf{z},\mathbf{t}} \mathbf{t}^T \mathbf{A}^T \text{diag}(\mathbf{u})\mathbf{c}$$

$$\text{s.t.} \begin{cases} (\mathbf{z} - \mathbf{A}\mathbf{x} - \mathbf{b})^2 = 0 \\ \mathbf{t}^2 \leq 1, (\mathbf{x} - \bar{\mathbf{x}}_0 + \varepsilon)(\mathbf{x} - \bar{\mathbf{x}}_0 - \varepsilon) \leq 0 \\ \mathbf{u}(\mathbf{u} - 1) = 0, (\mathbf{u} - 1/2)\mathbf{z} \geq 0 \end{cases}$$

- Pick SDP variables for products in $\{x, t\}$, $\{u, z\}$ up to deg 4
- Pick SDP variables for products in $\{x, z\}$, $\{t, u\}$ up to deg 2

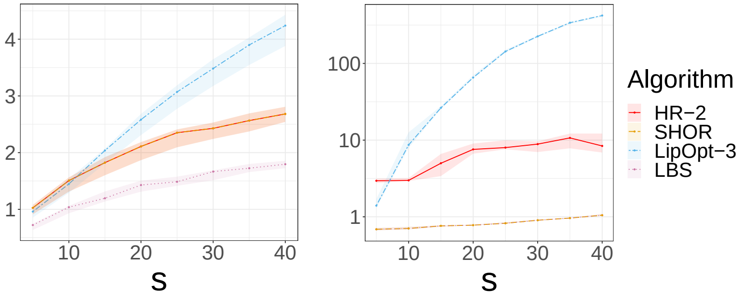
HR-2 on random $(80, 80)$ networks

Weight matrix $\mathbf{A}$ with band structure of width **s**

SHOR: Shor's relaxation given by 1ST stair in the hierarchy

LipOpt-3: LP based method

LBS: lower bound given by 10^4 random samples



Upper bound Time

HR-2 on trained $(784, 500)$ network

MNIST classifier (**SDP-NN**) from Raghunathan et al. *Certified defenses against adversarial examples*, ICLR'18

		HR-2		SHOR	LipOpt-3	LBS
Global Lipschitz	Bound	14.56	$<$	17.85	Out of RAM	9.69
	Time	12246	$>$	2869	Out of RAM	-
Local Lipschitz	Bound	12.70	$<$	16.07	-	8.20
	Time	20596	$>$	4217	-	-

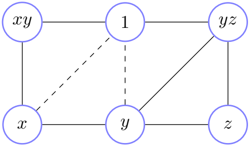
What's next?

MORE LAYERS $\Longrightarrow$ higher degree polynomials
TSSOS HIERARCHY: exploit term sparsity [Wang-M.-Lasserre 19]

What's next?

MORE LAYERS $\Longrightarrow$ higher degree polynomials
TSSOS HIERARCHY: exploit term sparsity [Wang-M.-Lasserre 19]

Term sparsity pattern graph
Chordal extension



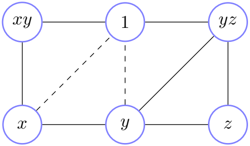
⇝ Link with Jared Miller's poster!

What's next?

MORE LAYERS $\Longrightarrow$ higher degree polynomials
TSSOS HIERARCHY: exploit term sparsity [Wang-M.-Lasserre 19]

💡 **Term** sparsity pattern graph
Chordal extension

$\rightsquigarrow$ Link with Jared Miller's poster!



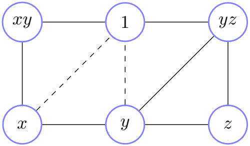
CERTIFIED bounds $\rightsquigarrow$ embed ML into "CRITICAL" dynamical systems

What's next?

MORE LAYERS $\implies$ higher degree polynomials
TSSOS HIERARCHY: exploit term sparsity [Wang-M.-Lasserre 19]

💡 **Term** sparsity pattern graph
Chordal extension

$\rightsquigarrow$ Link with Jared Miller's poster!



CERTIFIED bounds $\rightsquigarrow$ embed ML into "CRITICAL" dynamical systems

Open PhD/Postdoc positions

ANITI
ARTIFICIAL & NATURAL INTELLIGENCE TOULOUSE INSTITUTE
Université Fédérale Toulouse Midi-Pyrénées

Thank you for your attention!

`https://homepages.laas.fr/vmagron`

Chen, Lasserre, Magron and Pauwels. *Polynomial Optimization for Bounding Lipschitz Constants of Deep Networks*. arxiv:2002.03657

Wang, Magron & Lasserre. TSSOS: a moment-SOS hierarchy that exploits term sparsity. arxiv:1912.08899 TSSOS